US 20220383589A1

(19) **United States**

(12) **Patent Application Publication**
Fujita et al.

(10) **Pub. No.: US 2022/0383589 A1**

(43) **Pub. Date: Dec. 1, 2022**

(54) **IMAGE PROCESSING APPARATUS, IMAGE PROCESSING METHOD, METHOD FOR GENERATING LEARNED MODEL, AND STORAGE MEDIUM**

(71) Applicant: **CANON KABUSHIKI KAISHA**, Tokyo (JP)

(72) Inventors: **Shu Fujita**, Kanagawa (JP); **Keigo Yoneda**, Kanagawa (JP); **Shuntaro Aratani**, Tokyo (JP); **Atsushi Date**, Tokyo (JP); **Toshiaki Fujii**, Aichi (JP); **Keita Takahashi**, Aichi (JP); **Takashi Sugie**, Aichi (JP)

(21) Appl. No.: **17/819,095**

(22) Filed: **Aug. 11, 2022**

Related U.S. Application Data

(63) Continuation of application No. PCT/JP2021/003988, filed on Feb. 3, 2021.

(30) **Foreign Application Priority Data**

Feb. 14, 2020 (JP) 2020-023374

Publication Classification

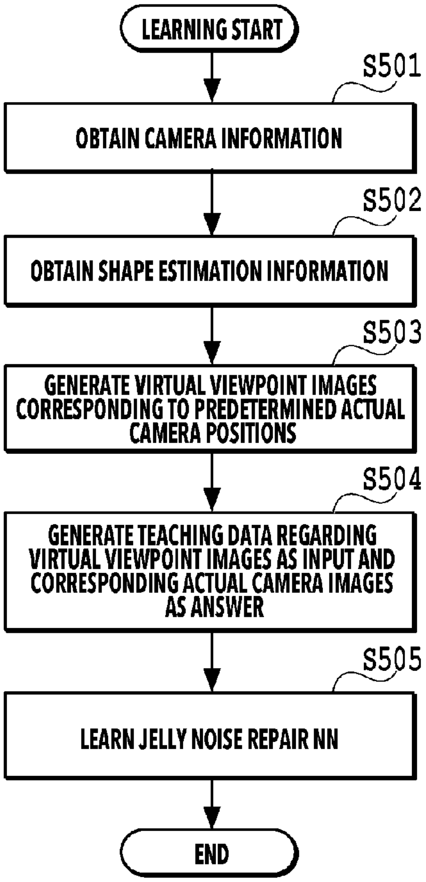
(51) **Int. Cl.**

G06T 15/20	(2006.01)
G06V 20/64	(2006.01)
G06T 5/00	(2006.01)
G06T 11/00	(2006.01)

(52) **U.S. Cl.**
CPC ***G06T 15/205*** (2013.01); ***G06V 20/647*** (2022.01); ***G06T 5/002*** (2013.01); ***G06T 11/001*** (2013.01); *G06T 2207/20081* (2013.01)

(57) **ABSTRACT**

An image generation apparatus obtains a virtual viewpoint image generated based on captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints and three-dimensional shape data on the object, and removes noise in the virtual viewpoint image obtained, the noise being generated due to accuracy of the three-dimensional shape data.


LEARNING START
S501
OBTAIN CAMERA INFORMATION
S502
OBTAIN SHAPE ESTIMATION INFORMATION
S503
GENERATE VIRTUAL VIEWPOINT IMAGES CORRESPONDING TO PREDETERMINED ACTUAL CAMERA POSITIONS
S504
GENERATE TEACHING DATA REGARDING VIRTUAL VIEWPOINT IMAGES AS INPUT AND CORRESPONDING ACTUAL CAMERA IMAGES AS ANSWER
S505
LEARN JELLY NOISE REPAIR NN
END

LEARNING

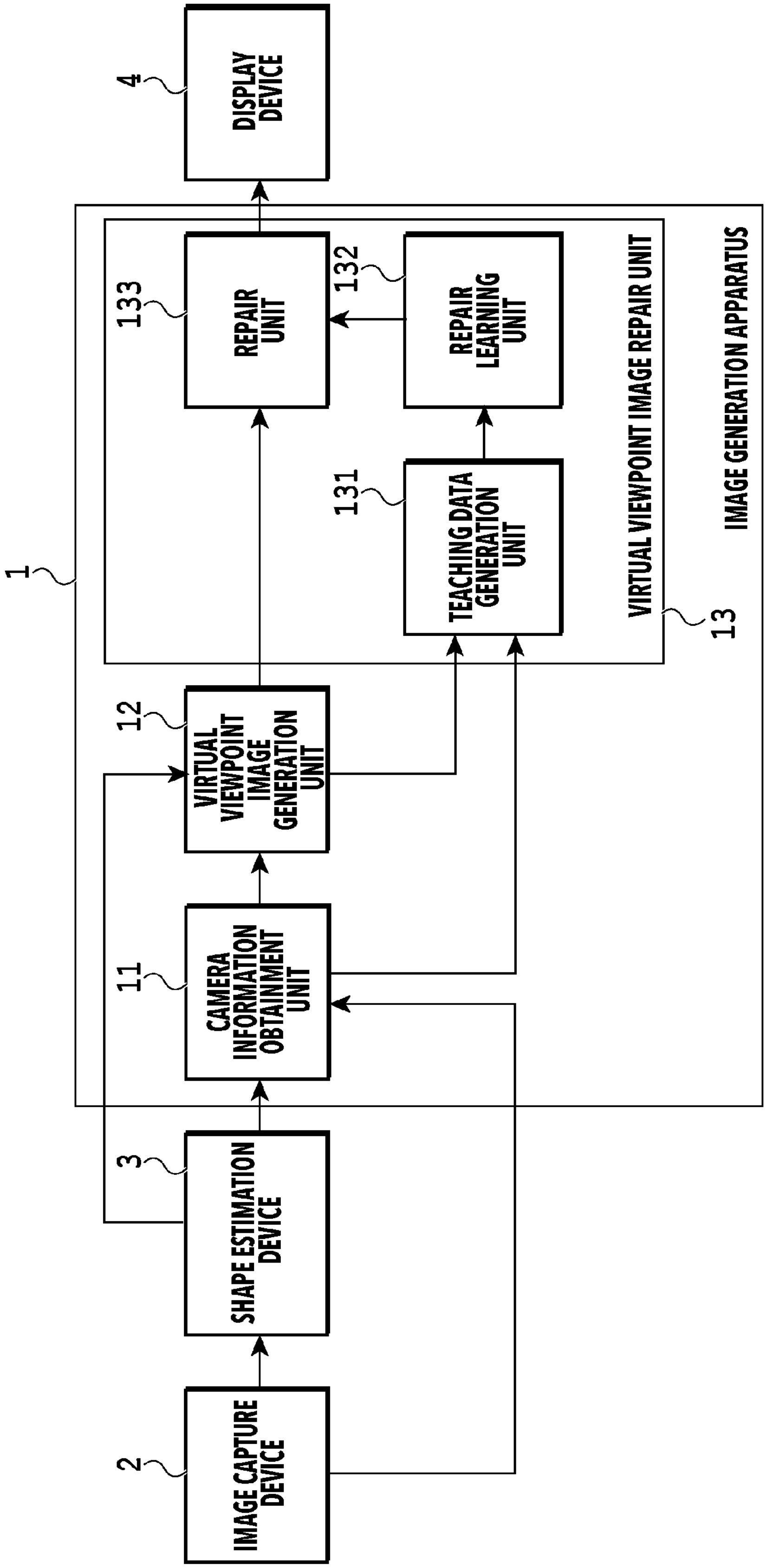

2
IMAGE CAPTURE DEVICE
3
SHAPE ESTIMATION DEVICE
1
11
CAMERA INFORMATION OBTAINMENT UNIT
12
VIRTUAL VIEWPOINT IMAGE GENERATION UNIT
13
131
TEACHING DATA GENERATION UNIT
132
REPAIR LEARNING UNIT
133
REPAIR UNIT
VIRTUAL VIEWPOINT IMAGE REPAIR UNIT
IMAGE GENERATION APPARATUS
4
DISPLAY DEVICE

FIG.1

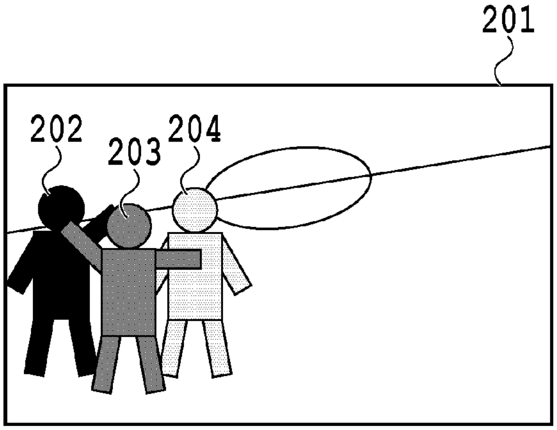


FIG.2A

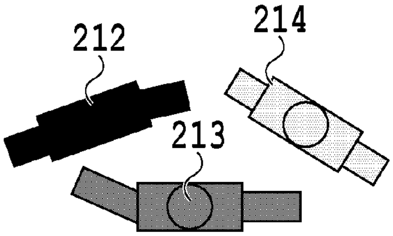


FIG.2B

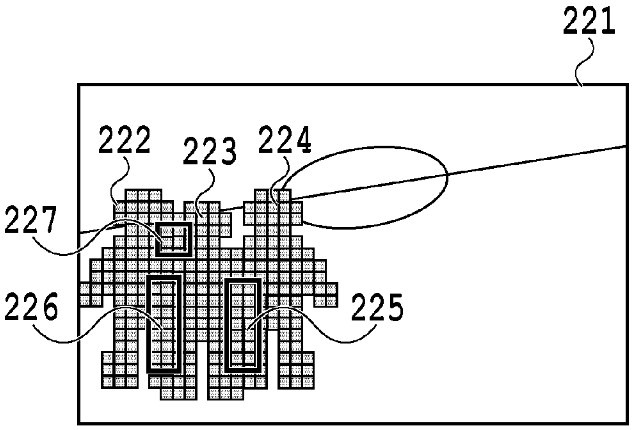


FIG.2C

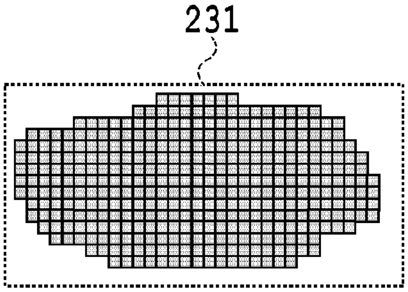


FIG.2D

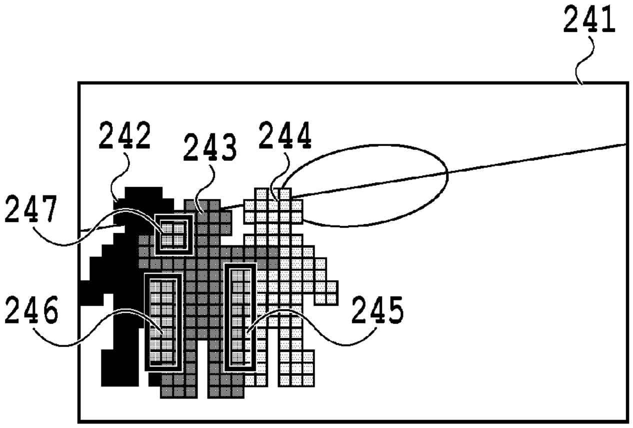


FIG.2E

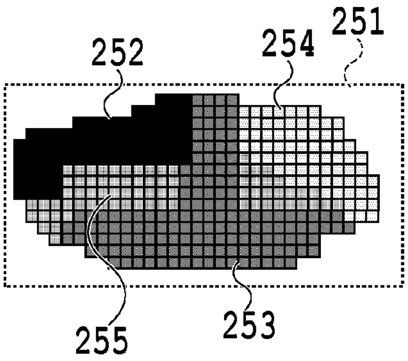


FIG.2F

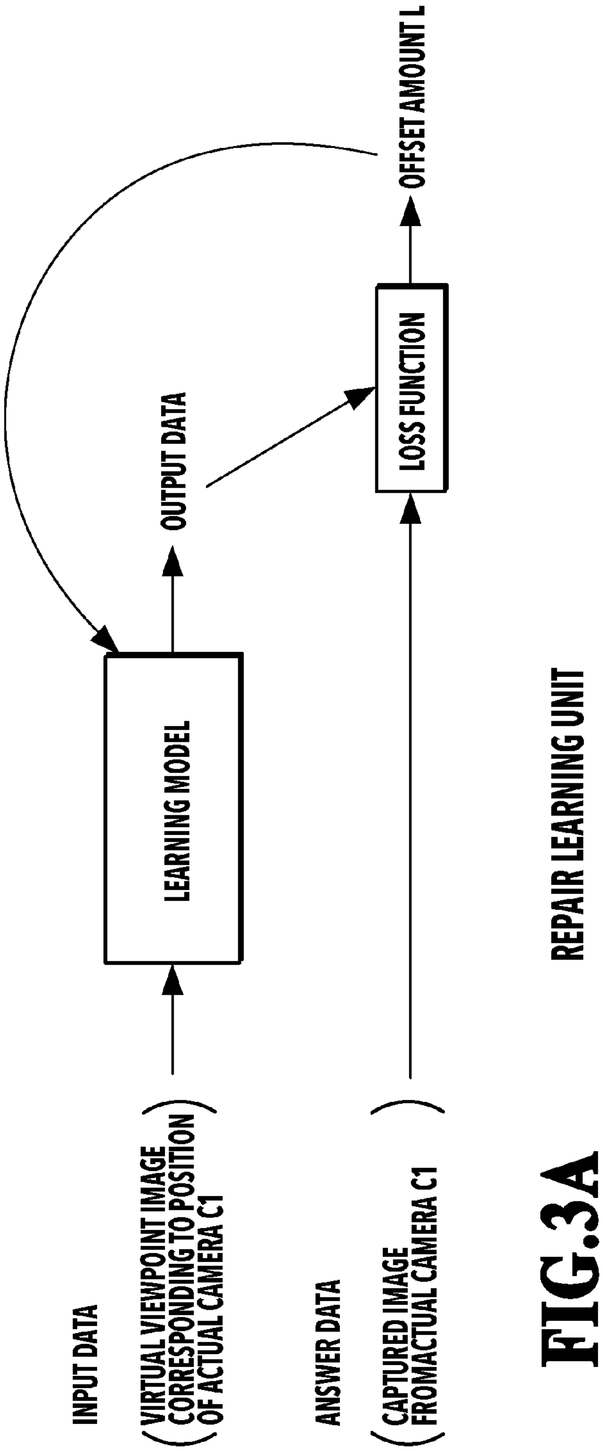

INPUT DATA
(VIRTUAL VIEWPOINT IMAGE CORRESPONDING TO POSITION OF ACTUAL CAMERA C1)
LEARNING MODEL
OUTPUT DATA
LOSS FUNCTION
OFFSET AMOUNT L
ANSWER DATA
(CAPTURED IMAGE FROMACTUAL CAMERA C1)
REPAIR LEARNING UNIT

FIG.3A

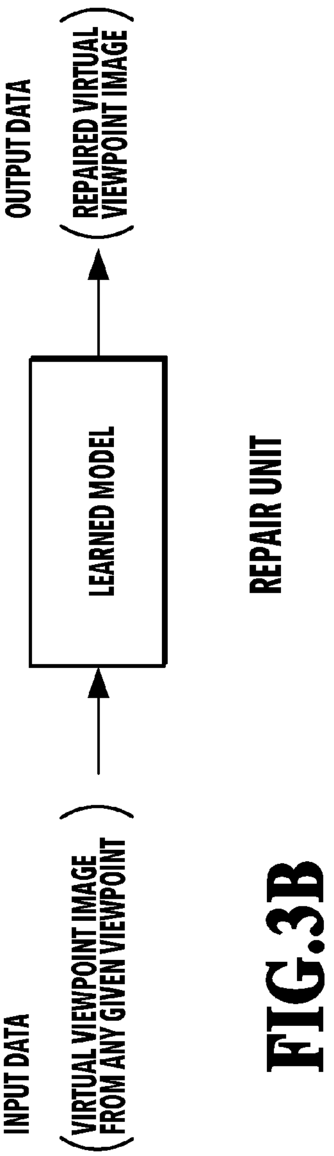

INPUT DATA
(VIRTUAL VIEWPOINT IMAGE FROM ANY GIVEN VIEWPOINT)
LEARNED MODEL
OUTPUT DATA
(REPAIRED VIRTUAL VIEWPOINT IMAGE)
REPAIR UNIT

FIG.3B

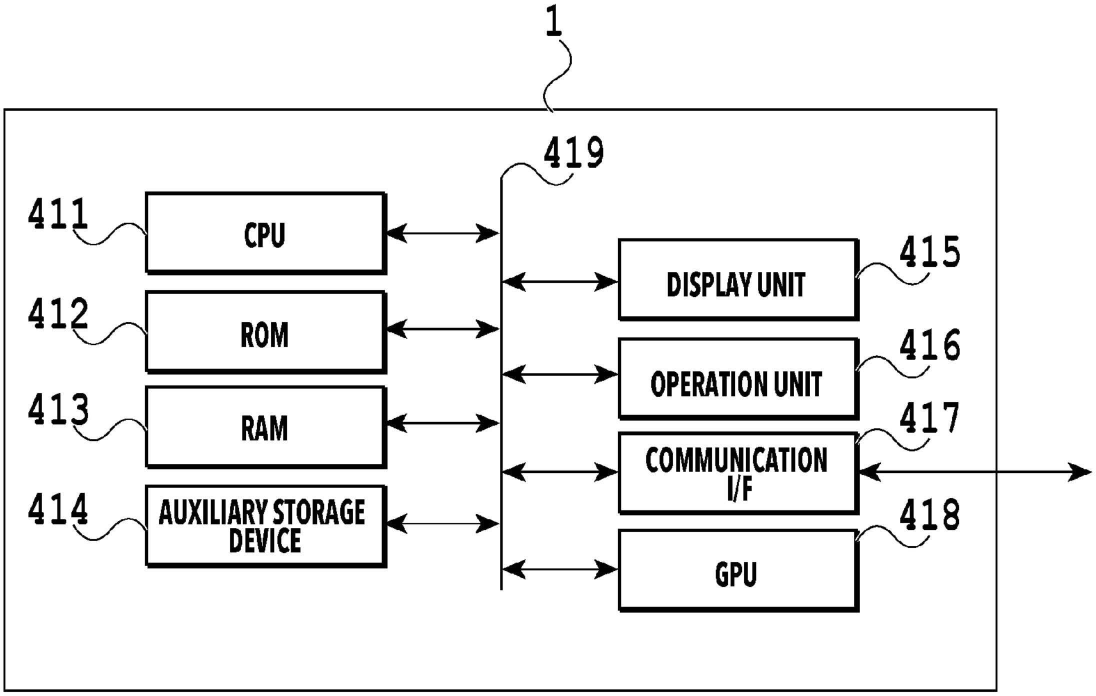

1
419
411
CPU
412
ROM
413
RAM
414
AUXILIARY STORAGE DEVICE
415
DISPLAY UNIT
416
OPERATION UNIT
417
COMMUNICATION I/F
418
GPU

FIG.4

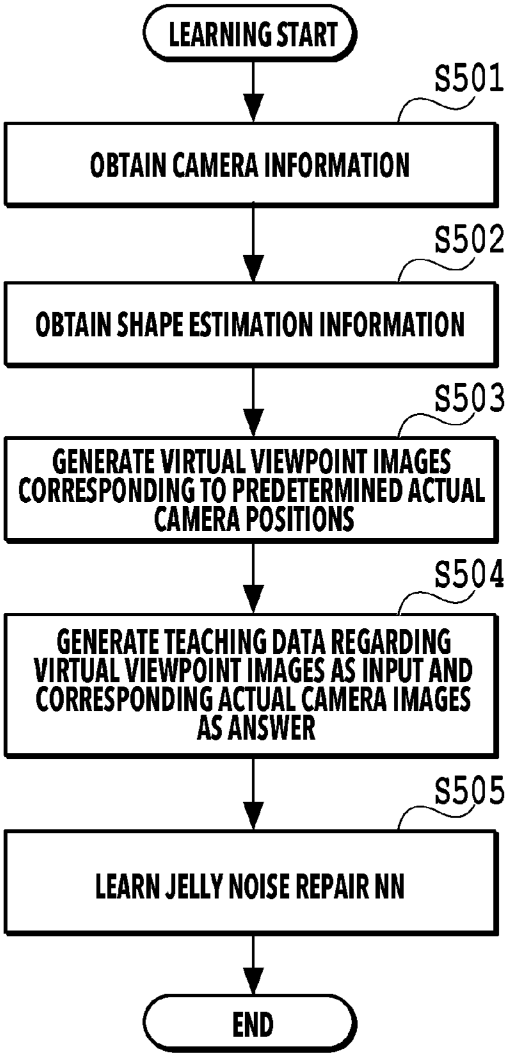

LEARNING START
S501
OBTAIN CAMERA INFORMATION
S502
OBTAIN SHAPE ESTIMATION INFORMATION
S503
GENERATE VIRTUAL VIEWPOINT IMAGES CORRESPONDING TO PREDETERMINED ACTUAL CAMERA POSITIONS
S504
GENERATE TEACHING DATA REGARDING VIRTUAL VIEWPOINT IMAGES AS INPUT AND CORRESPONDING ACTUAL CAMERA IMAGES AS ANSWER
S505
LEARN JELLY NOISE REPAIR NN
END

FIG.5A

LEARNING

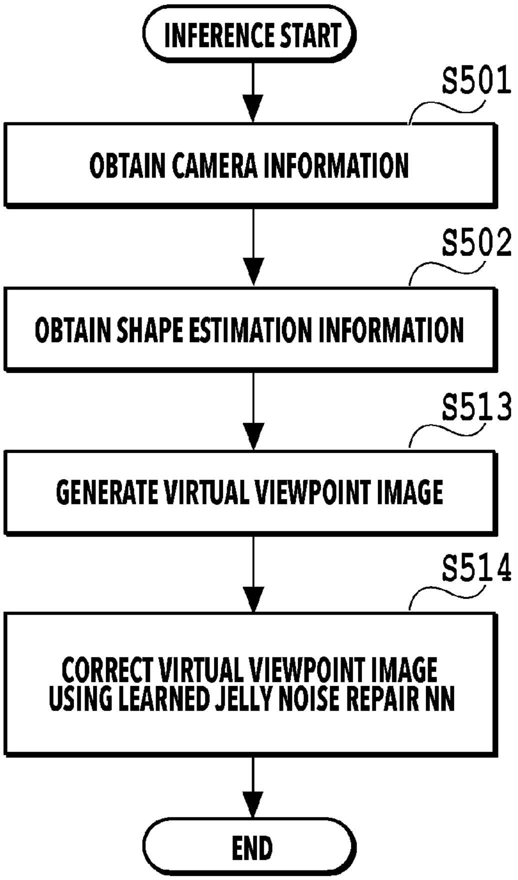

INFERENCE START
S501
OBTAIN CAMERA INFORMATION
S502
OBTAIN SHAPE ESTIMATION INFORMATION
S513
GENERATE VIRTUAL VIEWPOINT IMAGE
S514
CORRECT VIRTUAL VIEWPOINT IMAGE USING LEARNED JELLY NOISE REPAIR NN
END

FIG.5B

INFERENCE

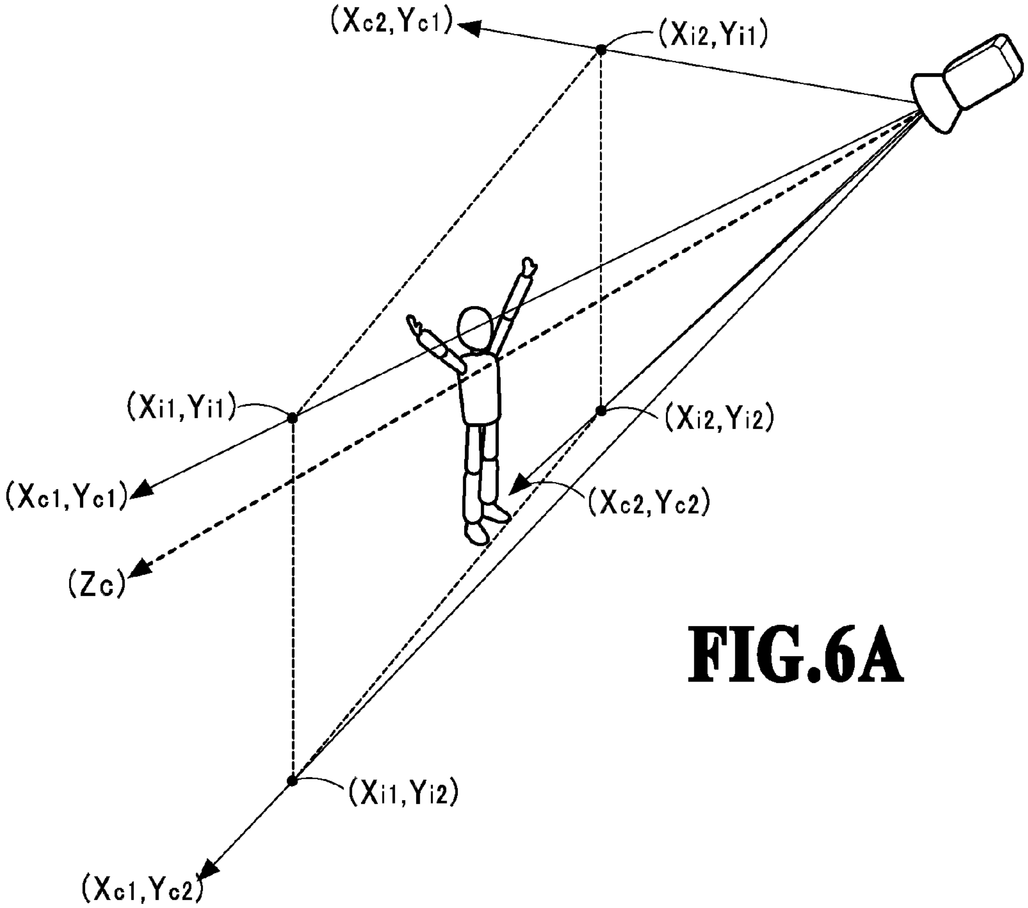
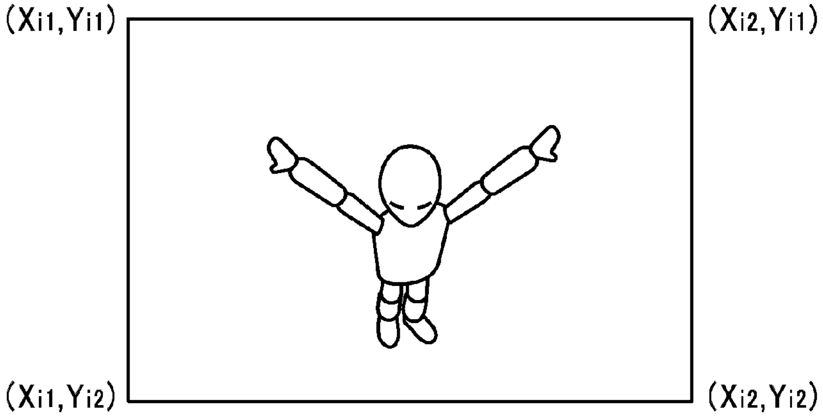


FIG.6A



CAMERA IMAGE COORDINATE SYSTEM

FIG.6B

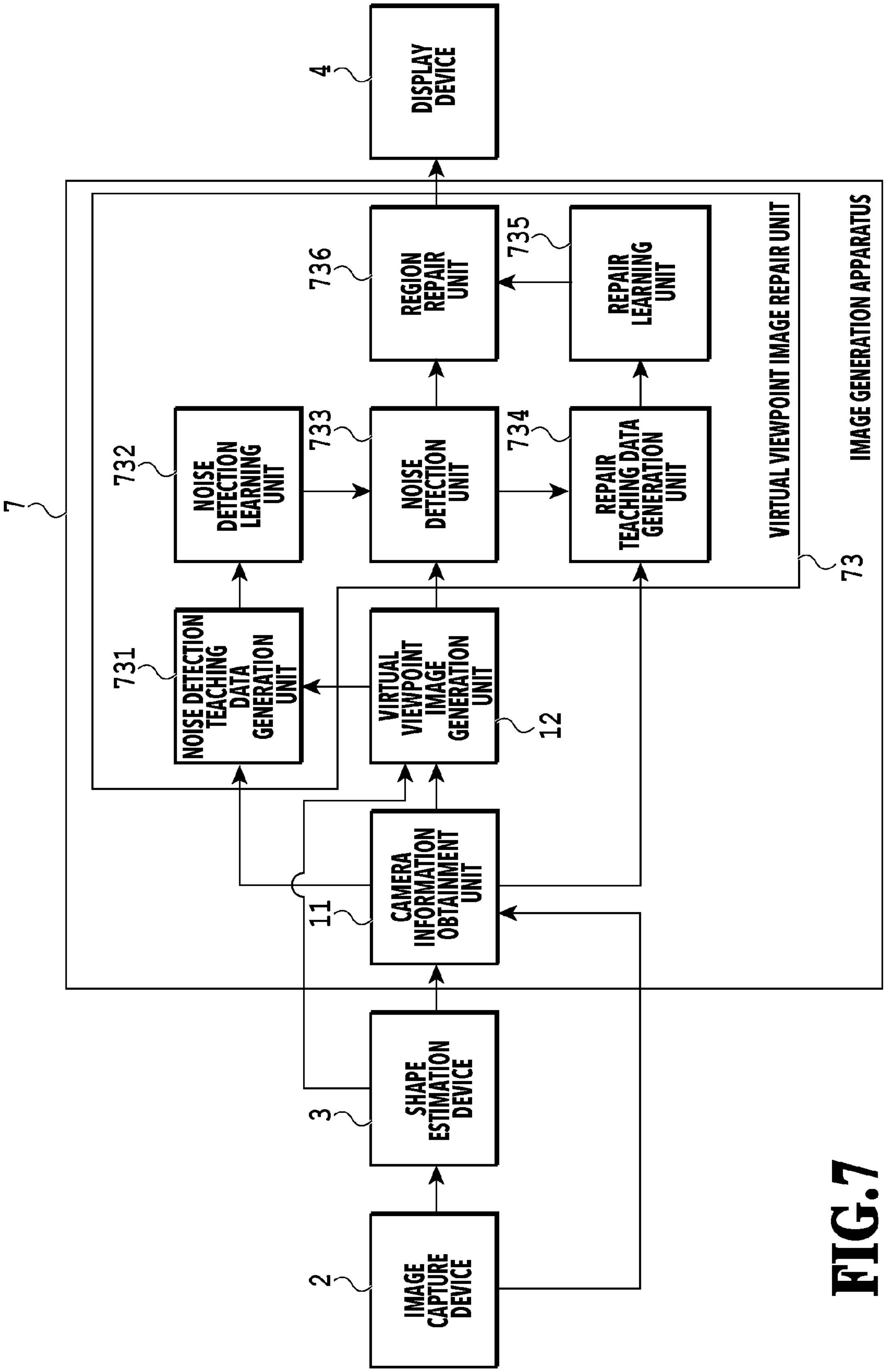

2
IMAGE CAPTURE DEVICE
3
SHAPE ESTIMATION DEVICE
7
IMAGE GENERATION APPARATUS
11
CAMERA INFORMATION OBTAINMENT UNIT
12
VIRTUAL VIEWPOINT IMAGE GENERATION UNIT
73
VIRTUAL VIEWPOINT IMAGE REPAIR UNIT
731
NOISE DETECTION TEACHING DATA GENERATION UNIT
732
NOISE DETECTION LEARNING UNIT
733
NOISE DETECTION UNIT
734
REPAIR TEACHING DATA GENERATION UNIT
735
REPAIR LEARNING UNIT
736
REGION REPAIR UNIT
4
DISPLAY DEVICE

FIG.7

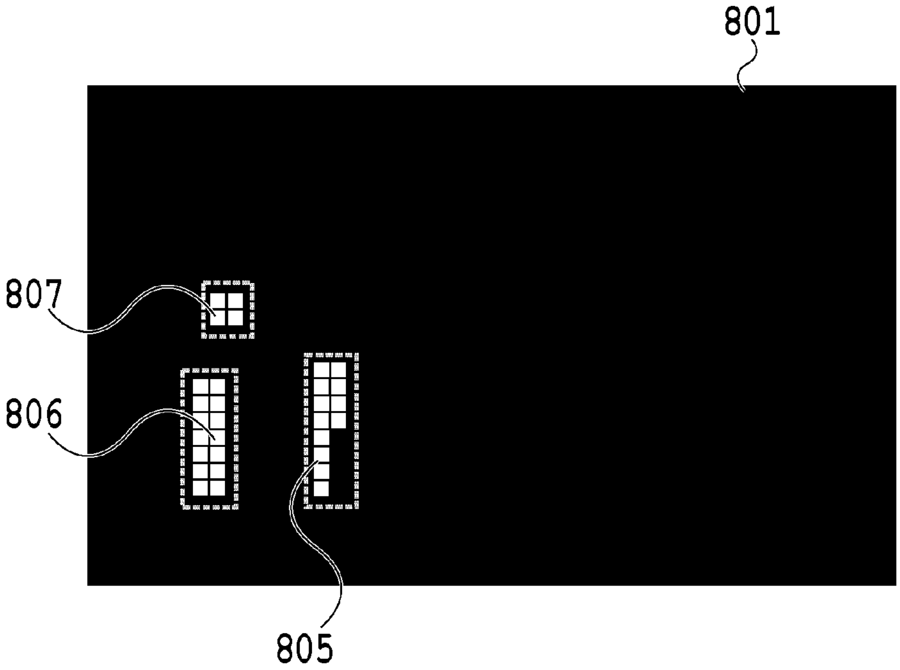


FIG.8A

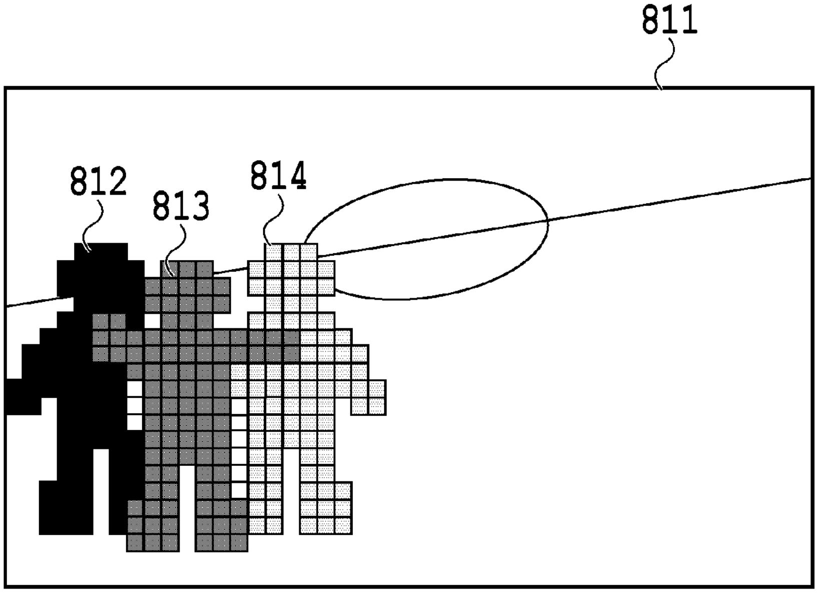


FIG.8B

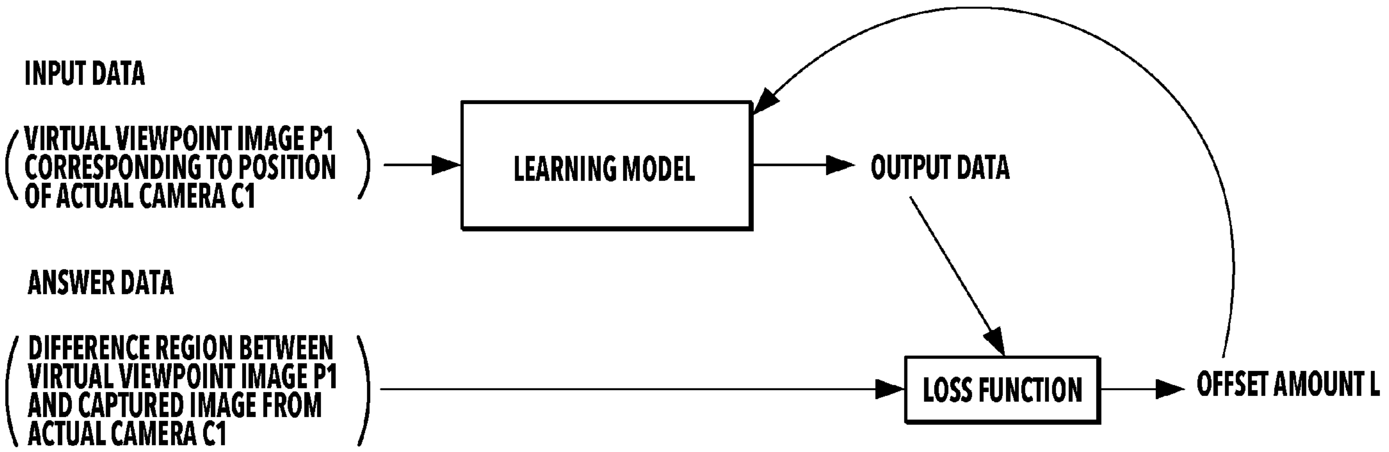

INPUT DATA
(VIRTUAL VIEWPOINT IMAGE P1 CORRESPONDING TO POSITION OF ACTUAL CAMERA C1)
LEARNING MODEL
OUTPUT DATA
ANSWER DATA
(DIFFERENCE REGION BETWEEN VIRTUAL VIEWPOINT IMAGE P1 AND CAPTURED IMAGE FROM ACTUAL CAMERA C1)
LOSS FUNCTION
OFFSET AMOUNT L

FIG.9A

NOISE DETECTION LEARNING UNIT

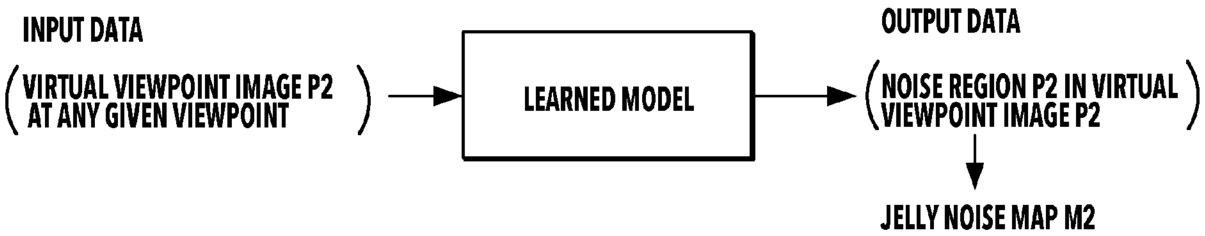

INPUT DATA
(VIRTUAL VIEWPOINT IMAGE P2 AT ANY GIVEN VIEWPOINT)
LEARNED MODEL
OUTPUT DATA
(NOISE REGION P2 IN VIRTUAL VIEWPOINT IMAGE P2)
JELLY NOISE MAP M2

FIG.9B

NOISE DETECTING UNIT

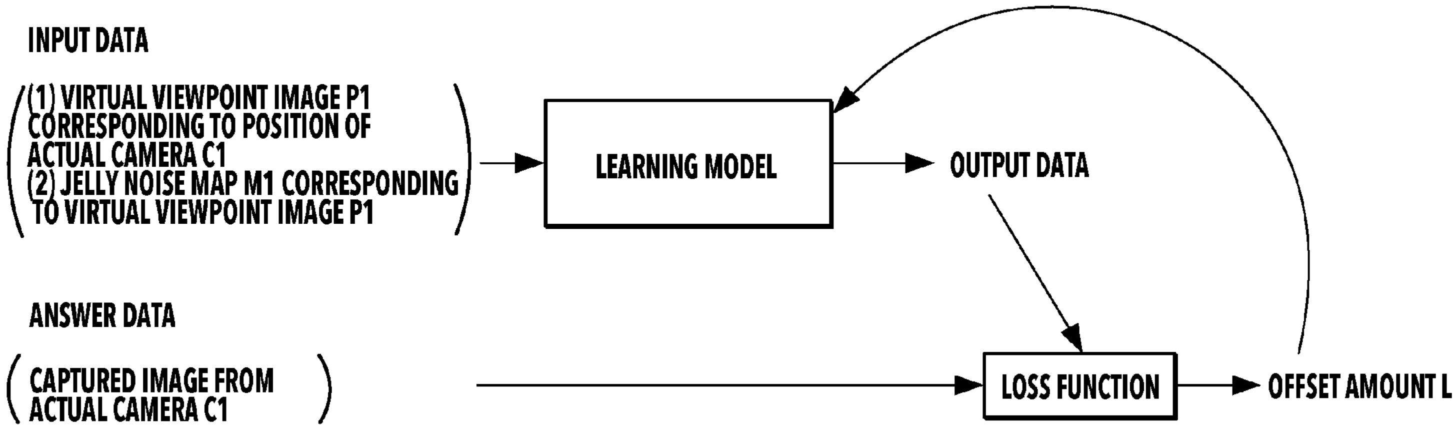

INPUT DATA
(1) VIRTUAL VIEWPOINT IMAGE P1 CORRESPONDING TO POSITION OF ACTUAL CAMERA C1
(2) JELLY NOISE MAP M1 CORRESPONDING TO VIRTUAL VIEWPOINT IMAGE P1
LEARNING MODEL
OUTPUT DATA
ANSWER DATA
CAPTURED IMAGE FROM ACTUAL CAMERA C1
LOSS FUNCTION
OFFSET AMOUNT L

FIG.10A

REPAIR LEARNING UNIT

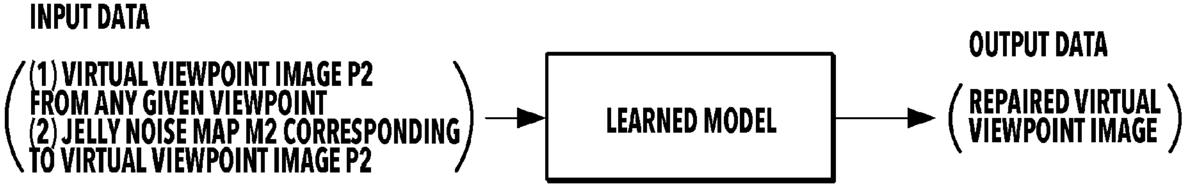

INPUT DATA
(1) VIRTUAL VIEWPOINT IMAGE P2 FROM ANY GIVEN VIEWPOINT
(2) JELLY NOISE MAP M2 CORRESPONDING TO VIRTUAL VIEWPOINT IMAGE P2
LEARNED MODEL
OUTPUT DATA
REPAIRED VIRTUAL VIEWPOINT IMAGE

FIG.10B

REGION REPAIR UNIT

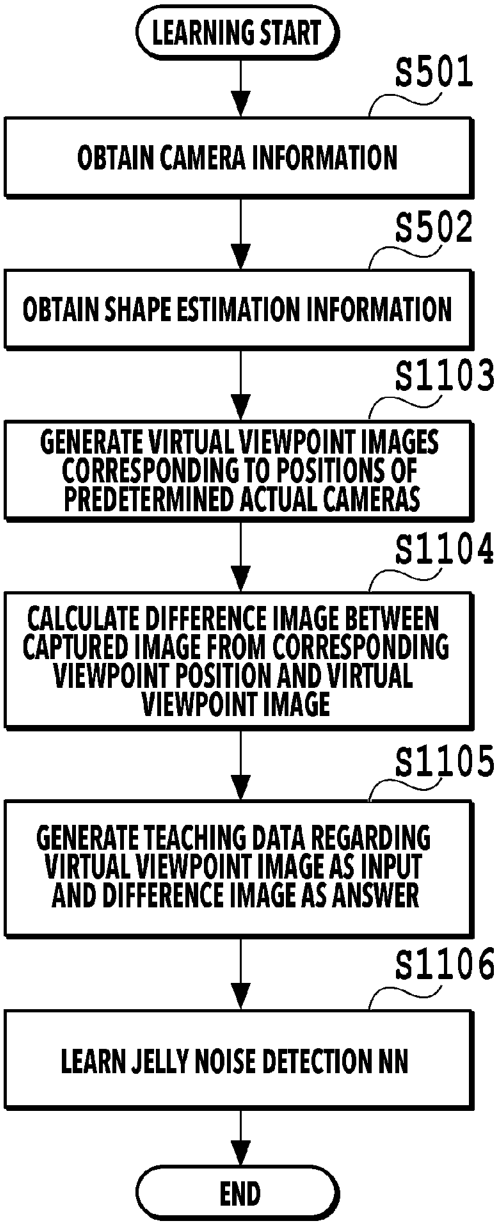

LEARNING START
S501
OBTAIN CAMERA INFORMATION
S502
OBTAIN SHAPE ESTIMATION INFORMATION
S1103
GENERATE VIRTUAL VIEWPOINT IMAGES CORRESPONDING TO POSITIONS OF PREDETERMINED ACTUAL CAMERAS
S1104
CALCULATE DIFFERENCE IMAGE BETWEEN CAPTURED IMAGE FROM CORRESPONDING VIEWPOINT POSITION AND VIRTUAL VIEWPOINT IMAGE
S1105
GENERATE TEACHING DATA REGARDING VIRTUAL VIEWPOINT IMAGE AS INPUT AND DIFFERENCE IMAGE AS ANSWER
S1106
LEARN JELLY NOISE DETECTION NN
END

FIG.11A

LEARNING FOR DETECTION

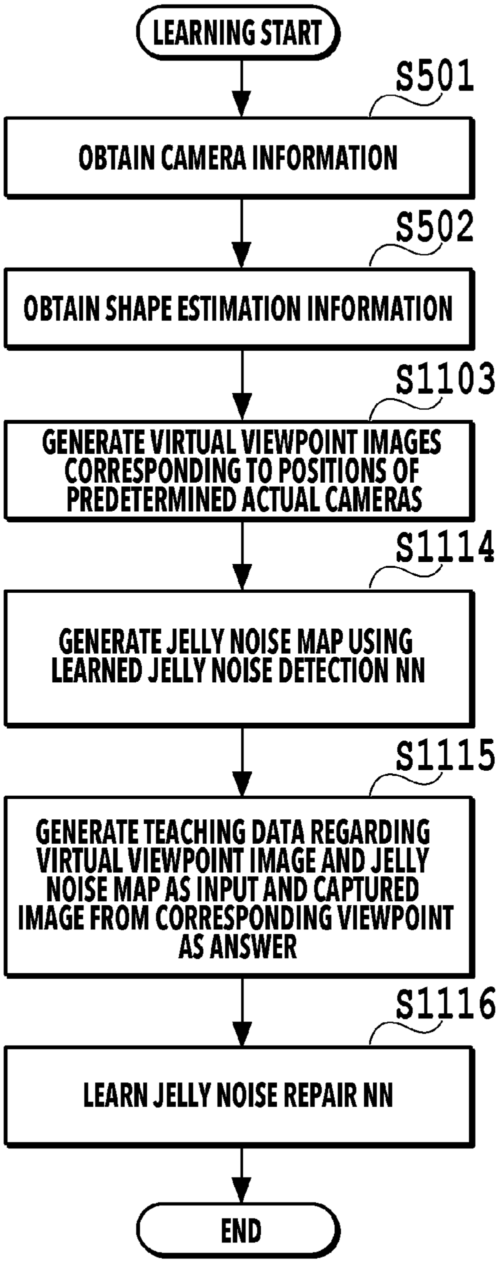

LEARNING START
S501
OBTAIN CAMERA INFORMATION
S502
OBTAIN SHAPE ESTIMATION INFORMATION
S1103
GENERATE VIRTUAL VIEWPOINT IMAGES CORRESPONDING TO POSITIONS OF PREDETERMINED ACTUAL CAMERAS
S1114
GENERATE JELLY NOISE MAP USING LEARNED JELLY NOISE DETECTION NN
S1115
GENERATE TEACHING DATA REGARDING VIRTUAL VIEWPOINT IMAGE AND JELLY NOISE MAP AS INPUT AND CAPTURED IMAGE FROM CORRESPONDING VIEWPOINT AS ANSWER
S1116
LEARN JELLY NOISE REPAIR NN
END

FIG.11B

LEARNING FOR REPAIR

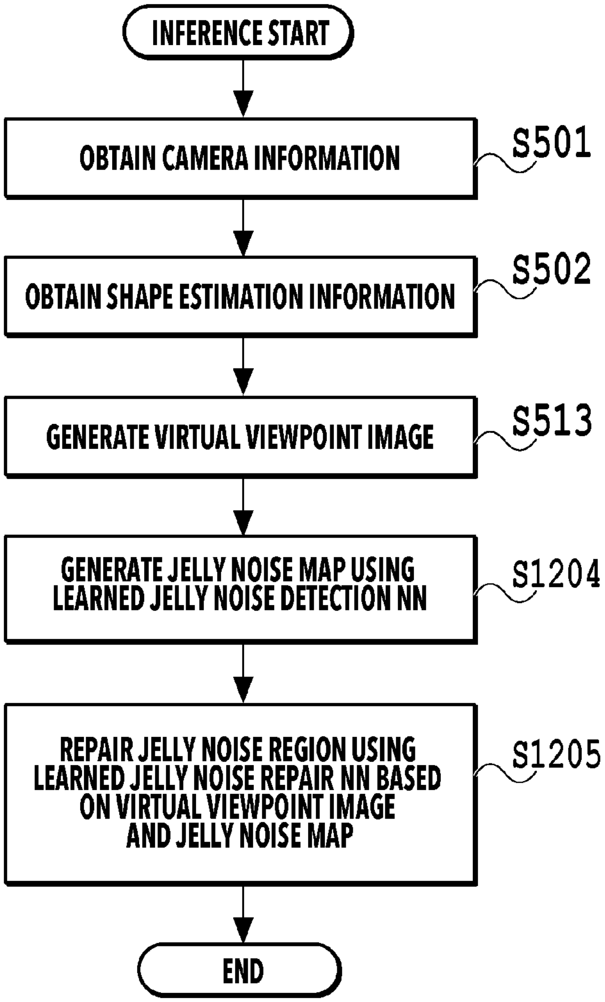

INFERENCE START
OBTAIN CAMERA INFORMATION
S501
OBTAIN SHAPE ESTIMATION INFORMATION
S502
GENERATE VIRTUAL VIEWPOINT IMAGE
S513
GENERATE JELLY NOISE MAP USING LEARNED JELLY NOISE DETECTION NN
S1204
REPAIR JELLY NOISE REGION USING LEARNED JELLY NOISE REPAIR NN BASED ON VIRTUAL VIEWPOINT IMAGE AND JELLY NOISE MAP
S1205
END

FIG.12

IMAGE PROCESSING APPARATUS, IMAGE PROCESSING METHOD, METHOD FOR GENERATING LEARNED MODEL, AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is a Continuation of International Patent Application No. PCT/JP2021/003988, filed Feb. 3, 2021, which claims the benefit of Japanese Patent Application No. 2020-023374, filed Feb. 14, 2020, both of which are hereby incorporated by reference herein in their entirety.

BACKGROUND

Field

[0002] The present disclosure relates to a virtual viewpoint image.

Background Art

[0003] There is technology for generating virtual viewpoint content representing a view from a virtual viewpoint using a plurality of images obtained by a plurality of image capture devices. Japanese Patent Laid-Open No. 2019-057248 (PTL 1) discloses generating virtual viewpoint content by first determining a color for each element forming a subject's three-dimensional shape estimated based on images obtained by image capture of the subject from a plurality of directions, the color being determined using the plurality of captured images.

CITATION LIST

Patent Literature

[0004] PTL 1 Japanese Patent Laid-Open No. 2019-057248

[0005] In a case where a virtual viewpoint image is generated by the method of PTL 1, the accuracy of the three-dimensional shape estimation affects the image quality of the virtual viewpoint image. In other words, without proper three-dimensional shape estimation, the image quality of the virtual viewpoint image may be degraded. For example, in a region in an image capture region where objects (subjects) are very close to each other, i.e., such as a region where occlusion occurs, an object which does not actually exist may be regarded as existing, and three-dimensional shape estimation may be performed thereon. In this case, among the plurality of elements forming the three-dimensional shape, incorrect colors are determined for elements of an object which does not actually exist but is determined as existing. As a result, noise may occur in the virtual viewpoint image, degrading its image quality.

SUMMARY

[0006] An image processing apparatus according to an aspect of the present disclosure is an image processing apparatus including: obtainment means for obtaining a virtual viewpoint image generated based on a plurality of captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints and three-dimensional shape data on the object; and removal means for removing noise in the virtual viewpoint image obtained by the obtainment means, the noise being generated due to accuracy of the three-dimensional shape data.

[0007] Further features of the present disclosure will become apparent from the following description of exemplary embodiments with reference to the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. **1** is a diagram showing an example configuration of an image processing system.

[0009] FIGS. **2**A to **2**F are diagrams illustrating an example of a case where jelly noise occurs.

[0010] FIGS. **3**A and **3**B are diagrams illustrating an overview of a learning model.

[0011] FIG. **4** is a diagram showing an example hardware configuration of an image generation apparatus.

[0012] FIGS. **5**A and **5**B are flowcharts showing an example of processing performed by the image generation apparatus.

[0013] FIGS. **6**A and **6**B are schematic diagrams of a camera coordinate system and a camera image coordinate system.

[0014] FIG. **7** is a diagram showing the configuration of the image processing system.

[0015] FIGS. **8**A and **8**B are diagrams illustrating a jelly noise map.

[0016] FIGS. **9**A and **9**B are diagrams illustrating an overview of a learning mode for detecting a jelly noise region.

[0017] FIGS. **10**A and **10**B are diagrams illustrating an overview of a learning model for repairing a jelly noise region.

[0018] FIGS. **11**A and **11**B are flowcharts showing an example of processing performed by an image generation apparatus **7**.

[0019] FIG. **12** is a flowchart showing an example of processing performed by the image generation apparatus **7**.

DESCRIPTION OF THE EMBODIMENTS

[0020] Modes for carrying out an aspect of the present disclosure are described below with reference to the drawings. Note that the following embodiments are not intended to limit the matters of the present disclosure, and not all the combinations of the features described in the present embodiments are necessarily essential to the solutions provided by an aspect of the matters of the present disclosure. Note that the same reference numeral is used to describe the same configurations.

First Embodiment

[0021] In the present embodiment, an example is discussed of performing processing for repairing (or mending or correcting) a virtual viewpoint image containing noise that occurs due to a result of low-accuracy shape estimation (hereinafter referred to as jelly noise). For the repairing processing, a learned model (a neural network (called an NN below)) is used. Specifically, as a result of inputting a virtual viewpoint image containing jelly noise to a learned model, a virtual viewpoint image removed of (improved in) a jelly noise part is outputted from the learned model.

[0022] Note that jelly noise occurs due to three-dimensional shape estimation estimating, because of occlusion, that an object (which may also be called a subject) which actually does not exist exists. Jelly noise is also likely to occur in an object having a complicated shape, such as one including many irregularities.

[0023] An image processing system of the present embodiment generates a virtual viewpoint image representing a view from a virtual viewpoint based on a plurality of captured images captured and obtained by a plurality of image capture devices from different directions, the states of the image capture devices, and virtual viewpoint information indicating the virtual viewpoint.

[0024] The plurality of image capture devices capture images of an image capture region from a plurality of different directions. The image capture region is, for example, a region surrounded by a plane and any given height in a stadium in which, e.g., rugby or soccer games are held. The plurality of image capture devices are installed at different locations and in different directions in such a manner as to surround such an image capture region, and capture images synchronously. Note that the image capture devices do not have to be installed along the entire perimeter of the image capture region, and may be installed only at part of the image capture region due to, e.g., restrictions on installment locations. There is no limitation as to the number of the image capture devices, and in an example where the image capture region is a rugby stadium, approximately several tens to several hundreds of image capture devices may be installed around the stadium.

[0025] A plurality of image capture devices having different angles of view, such as telephoto cameras and wide-angle cameras, may be installed. For example, using telephoto cameras allows images of an object to be captured at a high resolution and therefore improves the resolution of a virtual viewpoint image generated. Also, for example, using wide-angle cameras can reduce the number of cameras because a wide range can be captured by a single camera. The image capture devices are synchronized based on information on a single time in real world, and a captured video has image capture time information added to each image frame.

[0026] Note that one image capture device may be formed by one camera or may be formed by a plurality of cameras. Also, an image capture device may include a device other than a camera.

[0027] The states of an image capture device are the image capture device's position, attitude (orientation and image capture direction), focal length, optical center, distortion, and the like. The position and attitude (orientation and image capture direction) of an image capture device may be controlled by the image capture device itself or by control of a panhead for controlling the position and attitude of the image capture device. Although data indicative of the states of an image capture device are referred to as camera parameters of the image capture device in the following description, the parameters may include a parameter controlled by another device such as a panhead. Also, camera parameters related to the position and attitude (orientation and image capture direction) of an image capture device are what is called extrinsic parameters. Parameters related to the focal length, image center, and distortion of an image capture device are what is called intrinsic parameters. The position and attitude of an image capture device are expressed by a coordinate system having three axes orthogonal to a single origin (hereinafter referred to as a world coordinate system).

[0028] A virtual viewpoint image is also called a free viewpoint image, but a virtual viewpoint image is not limited to an image corresponding to a viewpoint designated freely (arbitrary) by a user, and includes, e.g., an image corresponding to a viewpoint selected by a user from a plurality of candidates. The designation of a virtual viewpoint may be performed by a user operation or automatically based on, e.g., image analysis results. Also, although a virtual viewpoint image is mainly described as being a still image in the present embodiment, a virtual viewpoint image may be a moving image.

[0029] Virtual viewpoint information used for generation of a virtual viewpoint image is information indicating, e.g., the position and orientation of a virtual viewpoint. Specifically, virtual viewpoint information includes parameters representing the three-dimensional position of a virtual viewpoint and parameters representing the orientation of the virtual viewpoint in pan, tilt, and roll directions. Note that the contents of the virtual viewpoint information are not limited to the above. For example, parameters in the virtual viewpoint information may include a parameter representing the size of the field of view (the angle of view) of the virtual viewpoint. Also, virtual viewpoint information may have parameters for a plurality of frames. Specifically, virtual viewpoint information may be information having parameters corresponding to a plurality of respective frames forming moving images of virtual viewpoint images and indicating the position and orientation of the virtual viewpoint at each of a plurality of consecutive time points.

[0030] For example, a virtual viewpoint image is generated by the following method. First, image capture devices capture their image capture regions from different directions, and a plurality of captured images are thereby obtained. Next, from the plurality of captured images, foreground images and background images are obtained, the foreground images being an extraction of a foreground region corresponding to an object such as a person or a ball, the background images being an extraction of a background region other than the foreground region. The foreground images and the background images have texture information (such as color information). Then, a foreground model representing the three-dimensional shape of the object and texture data for coloring the foreground model are generated based on the foreground images. The foreground model is estimated using a shape estimation method such as, for example, the Shape-from-Silhouette method. A background model is generated by making three-dimensional measurements of, for example, the stadium or venue in advance. Also, texture data for coloring a background model representing the three-dimensional shape of a background such as the stadium is generated based on the background images. Then, the texture data is mapped to the foreground model and the background model, and rendering is performed based on the virtual viewpoint indicated by the virtual viewpoint information, thereby generating a virtual viewpoint image. Note that the virtual viewpoint image generation method is not limited to this, and various methods can be used such as a method for generating a virtual viewpoint image by projective transformations of captured images, without using foreground models and background models.

[0031] A foreground image is an extracted image of the region of an object (a foreground region) from a captured image captured and obtained by an image capture device. An object extracted as a foreground region is typically a dynamic object (a dynamic body) which is active (may change in its position or shape) in a case where the object is captured chronologically from the same direction. Examples of an object include, in a sporting event, a person such as a player or a referee in the field where a game is held and may also include a ball in addition to a person in a case of a ball game. Also, in a case of a concert, an entertainment, or the like, examples of an object include a singer, a player, a performer, or an emcee.

[0032] A background image is an image of a region (a background region) different from at least a foreground object. Specifically, a background image is an image where foreground objects are removed from a captured image. Also, a background is an image capture target which is stationary or stays nearly stationary in a case where the background is captured chronologically from the same direction. Examples of such an image capture target include the stage for a concert or the like, a stadium where an event such as a sporting event is held, a structure such as a goal used in a ball game, and a field. Note, however, that a background is a region different from at least a foreground object, and an image capture target may also include physical objects and the like other than an object and a background.

<System Configuration>

[0033] FIG. **1** is a diagram showing an example configuration of the image processing system of the present embodiment. The image processing system has an image generation apparatus **1**, a plurality of image capture devices **2**, a shape estimation device **3**, and a display device **4**. FIG. **1** shows only one of the image capture devices **2**, omitting the rest of the image capture devices **2**.

[0034] The image generation apparatus **1** as an image processing apparatus is connected to the image capture devices **2**, the shape estimation device **3**, and the display device **4** in a daisy chain or via a predetermined network. The image generation apparatus **1** obtains captured image data from the image capture devices **2**. The image generation apparatus **1** also obtains object's three-dimensional shape data from the shape estimation device **3**. Then, the image generation apparatus **1** generates virtual viewpoint image data based on the captured image data obtained from the image capture devices **2** and the three-dimensional shape data obtained from the shape estimation device **3**. An image represented by captured image data is referred to as a captured image, and an image represented by virtual viewpoint image data is referred to as a virtual viewpoint image. To simplify descriptions, the following description may express, for example, obtaining or generating various kinds of image data simply as obtaining or generating a virtual viewpoint image. To generate a virtual viewpoint image, the image generation apparatus **1** receives designation of virtual viewpoint information and generates a virtual viewpoint image based on the virtual viewpoint information. For example, virtual viewpoint information is designated by a user (an operator) using an input unit (not shown) such as a joystick, a jog dial, a touch panel, a keyboard, and a mouse. Note that designation of virtual viewpoint information is not limited to this, and virtual viewpoint information may be designated automatically by, e.g., recognition of an object. A virtual viewpoint image generated by the image generation apparatus **1** is outputted to the display device **4**.

[0035] Each image capture device **2** has its own unique identification number so that the image capture device **2** may be distinguished from the other image capture devices **2**. The image capture device **2** may have other functions such as a function of extracting a foreground image from an image captured and obtained and may include hardware (such as a circuit or a device) for implementing that function.

[0036] The shape estimation device **3** obtains captured images or foreground images from the image capture devices **2**, estimates the three-dimensional shape of an object, and outputs three-dimensional shape data. The display device **4** obtains a virtual viewpoint image from the image generation apparatus **1** and outputs the virtual viewpoint image using a display device such as a display.

[0037] Next, the configuration of the image generation apparatus **1** is described. The image generation apparatus **1** has a camera information obtainment unit **11**, a virtual viewpoint image generation unit **12**, and a virtual viewpoint image repair unit **13**.

[0038] The camera information obtainment unit **11** obtains captured images from the plurality of image capture devices **2**. The camera information obtainment unit **11** also obtains camera parameters of each of the plurality of image capture devices **2**. Note that the camera information obtainment unit **11** may calculate and obtain the camera parameters of the image capture devices **2**. For example, the camera information obtainment unit **11** calculates corresponding points from the captured images obtained from the plurality of image capture devices **2**. Then, the camera information obtainment unit **11** calibrates the position, attitude, and the like of the viewpoint of each image capture device by performing optimization to minimize error in projection of the corresponding point to the viewpoint of the image capture device, and camera parameters may thus be calculated. The calibration method may be any of existing methods. Camera parameters may be obtained in synchronization with captured images, may be obtained in the preparation stage, or may be obtained out of synchronization with captured images as needed.

[0039] The virtual viewpoint image generation unit **12** generates a virtual viewpoint image based on captured images from the image capture devices **2** obtained by the camera information obtainment unit **11**, the camera parameters, three-dimensional shape data outputted from the shape estimation device **3**, and the virtual viewpoint information.

[0040] The virtual viewpoint image repair unit **13** repairs a virtual viewpoint image generated by the virtual viewpoint image generation unit **12**. This is because a virtual viewpoint image generated by the virtual viewpoint image generation unit **12** may contain jelly noise attributable to low-accuracy shape estimation. The virtual viewpoint image repair unit **13** removes this jelly noise.

<Description of Jelly Noise>

[0041] FIGS. **2**A to **2**F are diagrams illustrating an example of a case where the above-described jelly noise occurs due to low-accuracy shape estimation. Jelly noise is described using FIGS. **2**A to **2**F. FIG. **2**A shows a captured image **201** obtained by a certain image capture device **2** by capturing an image of objects. The captured image **201** shows objects **202**, **203**, **204**. FIG. **2**B shows an example of how the objects **202**, **203**, **204** look like from above. Objects **212**, **213**, **214** in FIG. **2**B correspond to the objects **202**, **203**, **204** in FIG. **2**A, respectively.

[0042] FIG. **2**C is an example of an image **221** of a case where a virtual viewpoint is designated at the viewpoint of a certain image capture device **2** that obtained the captured image **201**, using results of object shape estimation using the plurality of image capture devices **2** capturing the objects **202**, **203**, **204** in FIG. **2**A. Regions **222**, **223**, **224** in FIG. **2**C correspond to the objects **202**, **203**, **204**, respectively. Note that the colors have yet to be determined for the elements of the regions **222**, **223**, **224** in FIG. **2**C. FIG. **2**C shows that there are elements forming three-dimensional shape data corresponding to the regions **222**, **223**, **224**.

[0043] In a case where the objects **202**, **203**, **204** in FIG. **2**A are very close to each other, a region invisible from the plurality of image capture devices **2** occurs in the image capture region. In this case, the accuracy of shape estimation lowers, and for example, like regions **225**, **226**, **227** in FIG. **2**C, elements forming three-dimensional shape data may exist also in regions where objects do not actually exist. FIG. **2**D is a diagram showing three-dimensional shape data represented by the regions **222** to **227** in FIG. **2**C from above. In other words, as shown in FIG. **2**D, the regions **222** to **227** are formed as one lump of three-dimensional shape data **231** due to the influence of occlusion.

[0044] FIG. **2**E shows a virtual viewpoint image **241** obtained by coloring each element of the regions **222**, **223**, **224** in the image **221** in FIG. **2**C. Regions **242**, **243**, **244** in FIG. **2**E correspond to the regions **222**, **223**, **224** in FIG. **2**C, respectively. Regions **245**, **246**, **247** in FIG. **2**E correspond to the regions **225**, **226**, **227** in FIG. **2**C, respectively.

[0045] FIG. **2**F is a diagram showing three-dimensional shape data represented by the regions **242** to **247** in FIG. **2**E from above. Regions **252**, **253**, **254** in FIG. **2**F are three-dimensional shape data corresponding to the objects **212**, **213**, **214** in FIG. **2**B. It can be expected that three-dimensional points at positions where objects exist, like the regions **242**, **243**, **244** in FIG. **2**E or the regions **252**, **253**, **254** in FIG. **2**F, have the same colors as those of the original objects. However, for locations where objects do not actually exit, like the regions **245**, **246**, **247** in FIG. **2**E, it is highly likely that incorrect colors are assigned. A data region **255** which is part of the three-dimensional shape data in FIG. **2**F is a region corresponding to an occlusion region surrounded by the objects **212**, **213**, **214** in FIG. **2**B. As a result of incorrect colors being assigned to these locations, a virtual viewpoint image containing jelly noise is generated, which means the image quality of the virtual viewpoint image being low. This is an example of how jelly noise is generated.

<Description of the Virtual Viewpoint Image Repair Unit>

[0046] Referring back to FIG. **1**, the description is continued. The virtual viewpoint image repair unit **13** repairs a virtual viewpoint image generated by the virtual viewpoint image generation unit **12** that may contain jelly noise. Note that the present embodiment assumes that a learned model obtained by neural network learning is generated, and a virtual viewpoint image is repaired using the learned model. The following describes the virtual viewpoint image repair unit **13**. The virtual viewpoint image repair unit **13** has a teaching data generation unit **131**, a repair learning unit **132**, and a repair unit **133**.

[0047] The teaching data generation unit **131** generates teaching data having a pair of an input and an answer, the input being a virtual viewpoint image generated by the virtual viewpoint image generation unit **12**, the answer being a captured image from a camera having the corresponding viewpoint obtainable from the camera information obtainment unit **11**. Note that an image as answer data may be an image obtained by actually shooting a real space or an image generated by interpolation of captured images from two actual cameras. Also, an image as answer data may be an image obtained by combining captured images from three or more actual cameras. Also, a camera simulation image obtained in a virtual three-dimensional space created by CG (computer graphics) may be used. Note, however, that in a case where a camera used for actual shooting is used as answer data, the position and attitude of the virtual viewpoint of a virtual viewpoint image to be inputted are limited to the position and attitude of the actual camera. Also, in a case where an image generated by interpolation of captured images from two actual cameras is used as answer data, two cameras having their image capture regions overlapping with each other are selected, and only a region captured by both or one of the cameras is effective answer data. Also, in a case of using a CG simulation image as answer data, the correct three-dimensional shape of an object is already known. However, the virtual viewpoint image generation unit **12** does not use the correct three-dimensional shape. Instead, a plurality of pieces of camera information obtained by simulation are inputted to the shape estimation device **3**, and the virtual viewpoint image generation unit **12** uses, as an input, a virtual viewpoint image generated using a three-dimensional shape outputted from the shape estimation device **3**. Also in cases of using an image generated by interpolation of captured images from actual cameras or a CG simulation image as answer data, a viewpoint corresponding to these images is used as the viewpoint of a virtual viewpoint image used as an input. In other words, the teaching data generation unit **131** generates teaching data in which the position and attitude of the viewpoint of a virtual viewpoint image as an input corresponds to the position and attitude of an image as answer data. In this way, the teaching data generation unit **131** generates proper teaching data. Note that teaching data is also called learning data.

[0048] Based on the teaching data generated by the teaching data generation unit **131**, the repair learning unit **132** conducts learning by defining a loss function of the input with respect to the answer and repeatedly optimizing neural network parameters to minimize or maximize the loss function. Then, a model obtained by the learning (called a learned model) is outputted to the repair unit **133**.

[0049] FIGS. **3**A and **3**B are diagrams illustrating an overview of a learning model. FIG. **3**A shows an example of learning processing performed by the repair learning unit **132**. For example, learning is performed using teaching data having input data and answer data, the input data being a virtual viewpoint image corresponding to the viewpoint position of an actual camera C**1**, the answer data being a captured image captured by the actual camera C**1**. Then, learning is repeated to minimize or maximize an offset amount L between the input data and the answer data. Although an actual camera at one viewpoint is taken as an example here, learning is performed repeatedly using teaching data at the corresponding viewpoints of the image capture devices **2** forming the image processing system.

[0050] Note that the repair learning unit **132** may include an error detecting unit and an updating unit. The error detecting unit obtains error between teaching data and output data outputted from an output layer of a neural network in response to input data inputted to an input layer. The error detecting unit may calculate error between the teaching data and the output data from the neural network using a loss function. Based on the error obtained by the error detecting unit, the updating unit updates, e.g., connection weighting coefficients between nodes of the neural network so as to make the error small. The updating unit performs the update of the connection weighting coefficients or the like using, for example, backpropagation. Backpropagation is an algorithm for adjusting, e.g., a connection weighting coefficient between nodes of the neural network so as to make the above error small. Also, the present embodiment assumes that deep learning, which itself generates feature amounts and connection weighting coefficients for learning, is performed using a neural network. Note that as the network structure of a neural network used, any method may be employed as long as an input to and an output from the network are image data and the relation between the input and the output can be learned sufficiently.

[0051] The repair unit **133** repairs a virtual viewpoint image containing jelly noise by inputting the virtual viewpoint image given from the virtual viewpoint image generation unit **12** to the learned model obtained by the repair learning unit **132**. The repaired virtual viewpoint image is outputted to the display device **4**.

[0052] FIG. **3**B shows an example of repair processing (inference processing) performed by the repair unit **133**. In response to a virtual viewpoint image of any given virtual viewpoint being inputted as input data to the learned model obtained by the repair learning unit **132**, a repaired virtual viewpoint image is outputted as output data.

<Hardware Configuration>

[0053] FIG. **4** is a diagram showing an example hardware configuration of the image generation apparatus **1**. The image generation apparatus **1** has a CPU **411**, a ROM **412**, a RAM **413**, an auxiliary storage device **414**, a display unit **415**, an operation unit **416**, a communication I/F **417**, a GPU **418**, and a bus **419**. The CPU **411** implements the functions of the image generation apparatus **1** shown in FIG. **1** by performing overall control of the image generation apparatus **1** using computer programs and data stored in the ROM **412** or the RAM **413**. Note that the image generation apparatus **1** may have one or more pieces of dedicated hardware different from the CPU **411** and have the dedicated hardware execute at least part of the processing otherwise performed by the CPU **411**. Examples of the dedicated hardware include an ASIC (Application-Specific Integrated Circuit), an FPGA (Field-Programmable Gate Array), and a DSP (Digital Signal Processor). The ROM **412** stores programs and the like that do not need changes. The RAM **413** temporarily stores therein programs and data supplied from the auxiliary storage device **414** and data and the like supplied from outside via the communication I/F **417**. The auxiliary storage device **414** is formed of, for example, a hard disk drive or the like, and stores therein various kinds of data such as image data and audio data. The GPU **418** is capable of efficient computation by processing more pieces of data in parallel, and therefore it is effective to perform processing using the GPU **418** in a case of performing learning over a plurality of times using a learning model, such as deep learning. Thus, in addition to the CPU **411**, the GPU **418** is used for the processing by the repair learning unit **132** in the present embodiment. Specifically, in a case of executing a learning program including a learning model, learning is performed by the CPU **411** and the GPU **418** performing computations in cooperation with each other. Note that only one of the CPU **411** and the GPU **418** may perform computations for the processing by the repair learning unit **132**. Also, the repair unit **133** may use the GPU **418** like the repair learning unit **132** does.

[0054] The display unit **415** is formed of, for example, a liquid crystal display, an LED, or the like, and displays, e.g., a GUI (Graphical User Interface) for a user to operate the image generation apparatus **1**. The operation unit **416** is formed by, for example, a keyboard, a mouse, a joy stick, a touch panel, or the like, and inputs various instructions to the CPU **411** in response to user operations. The CPU **411** operates as a display control unit controlling the display unit **415** and as an operation control unit controlling the operation unit **416**.

[0055] The communication I/F **417** is used for communications between the image generation apparatus **1** and an external device. For example, in a case where the image generation apparatus **1** is connected to an external device in a wired manner, a communication cable is connected to the communication I/F **417**. In a case where the image generation apparatus **1** has a function of communicating wirelessly with an external device, the communication I/F **417** includes an antenna. The bus **419** connects the units in the image generation apparatus **1** to one another to communicate information thereamong.

[0056] The display unit **415** and the operation unit **416** are inside the image generation apparatus **1** in the present embodiment, but at least one of the display unit **415** and the operation unit **416** may be outside the image generation apparatus **1** as a separate device.

<Processing Flow>

[0057] FIGS. **5**A and **5**B are flowcharts showing an example of processing performed by the image generation apparatus **1** of the present embodiment. The processing shown in FIGS. **5**A and **5**B is performed by the CPU **411** or the GPU **418** executing programs stored in the ROM **412** or the auxiliary storage device **414**. Note that the letter “S” in the description of each processing means that it is a step in the flowchart (the same applies to the rest of the descriptions herein).

[0058] FIG. **5**A is a flowchart showing learning processing performed by the repair learning unit **132**. First, using FIG. **5**A, a description is given of a flowchart of processing for learning of a neural network for repairing a virtual viewpoint image.

[0059] In S**501**, the camera information obtainment unit **11** obtains camera information from the image capture devices **2**. Camera information may include a captured image and camera parameters. Specifically, in S**501**, the camera information obtainment unit **11** obtains a plurality of captured images from the image capture devices **2**. The captured images thus obtained are outputted to the virtual viewpoint image generation unit **12** and the teaching data generation unit **131**. Note that the captured images obtained here are used as answer data in neural network learning. In S**501**, the camera information obtainment unit **11** also obtains camera parameters from the image capture devices **2**. Note that the camera information obtainment unit **11** may calculate the camera parameters. Also, the camera parameters do not need to be calculated every time captured images are obtained, and only needs to be calculated at least once before generation of a virtual viewpoint image. The camera parameters thus obtained are outputted to the virtual viewpoint image generation unit **12**.

[0060] In S**502**, the virtual viewpoint image generation unit **12** obtains information on a group of three-dimensional points forming an object (three-dimensional shape data) from the shape estimation device **3**.

[0061] In S**503**, the virtual viewpoint image generation unit **12** generates a group of virtual viewpoint images corresponding to the positions of the viewpoints of the actual cameras. The group of virtual viewpoint images thus generated are outputted to the teaching data generation unit **131**. Thus, the virtual viewpoint images generated in S**503** are used as input data for neural network learning. In S**503** of this example, virtual viewpoint images corresponding to the viewpoint positions of all the actual cameras are generated. However, in this example, not all the frames of these virtual viewpoint images are outputted to the teaching data generation unit **131**, and a user selects in advance frames containing jelly noise and frames not containing jelly noise from frames shooting any foreground object in the virtual viewpoint images. Then, the virtual viewpoint image generation unit **12** outputs, to the teaching data generation unit **131**, virtual viewpoint images selected randomly so that there are an equal scene ratio of frames containing jelly noise and frames not containing jelly noise. By also including scenes in which no jelly noise occurs as learning input data, it is expected that a region that does not need a major correction unlike a jelly noise region can also be learned.

[0062] Here, a method for generating a virtual viewpoint image is described. The virtual viewpoint image generation unit **12** executes processing for generating a foreground virtual viewpoint image (a virtual viewpoint image of an object region) and processing for generating a background virtual viewpoint image (a virtual viewpoint image other than an object region). The virtual viewpoint image generation unit **12** then superimposes the foreground virtual viewpoint image onto the background virtual viewpoint image thus generated, thereby generating a virtual viewpoint image.

[0063] A method for generating a foreground virtual viewpoint image of a virtual viewpoint image is described. A foreground virtual viewpoint image can be generated by calculating the color of each voxel and rendering the colored voxel using an existing CG rendering method, assuming that each voxel is a three-dimensional point represented by coordinates (Xw, Yw, Zw). Before the color calculation, first, a distance image is generated in which each pixel value represents the distance from the camera of the image capture device **2** to the surface of the three-dimensional shape of an object.

[0064] A method for generating a distance image is described. A distance image has the same width and height as a captured image and has a distance value stored in each pixel. For example, an extrinsic matrix Te is applied to the coordinates (Xw, Yw, Zw) of a point P in a group of three-dimensional points to convert the coordinates from the coordinates of a world coordinate system to camera coordinates (Xc, Yc) of a camera coordinate system. A camera coordinate system is a three-dimensional coordinate system having the center of the camera lens as its origin and defined by a lens plane (Xc, Yc) and a lens optical axis (Zc). The extrinsic matrix Te is a matrix formed by extrinsic parameters of the actual camera. With the direction in which the actual camera lens is oriented from the camera position as the origin being the positive direction of the z-axis of the camera coordinate system, the z-coordinate of the camera coordinates (Xc, Yc) is a distance value for that point as seen from the actual camera.

[0065] Next, image coordinates (Xi, Yi) of the camera coordinates (Xc, Yc) are calculated, and coordinates in a distance image at which to store the distance value are obtained. The image coordinates (Xi, Yi) are coordinates in a camera image coordinate system calculated by applying an intrinsic matrix Ti to normalized camera coordinates obtained by normalization of the camera coordinates (Xc, Yc) with the z-coordinate. The camera image coordinate system is, as shown in FIG. **6**B, a two-dimensional coordinate system defined on a plane located forward of a lens surface by a certain distance and such that the Xc-axis and the Yc-axis of the camera coordinate system and the Xi-axis and the Yi-axis of the camera image coordinate system are parallel to each other, respectively. FIG. **6**A is a schematic diagram of the camera coordinate system, and FIG. **6**B is a schematic diagram of the camera image coordinate system. Note that the intrinsic matrix Ti is a matrix formed by intrinsic parameters of the actual camera. In a case where a distance value of another point already calculated is stored in the pixel of the image coordinates (Xi, Yi), this value is compared with the z-coordinate of the image coordinates (Xi, Yi). Then, in a case where the z-coordinate is smaller, the z-coordinate is stored anew as the pixel value of the image coordinates (Xi, Yi). By executing the series of processing on all the points P of the group of three-dimensional points, a distance image of a single actual camera can be generated. Further, by performing processing for all the actual cameras, distance images of all the actual cameras can be generated.

[0066] Next, to assign color to a voxel, with respect to a camera including a three-dimensional point (Xw, Yw, Zw) inside its angle of view, the three-dimensional point is first converted to the camera coordinate system. Then, the three-dimensional point thus converted to the camera coordinate system is converted to the camera image coordinate system, and a distance d from the voxel to the camera and coordinates (Xi, Yi) in the camera image coordinate system are calculated. Then, the difference between the distance d and the pixel value of the coordinates (Xi, Yi) corresponding to the distance image generated previously (=the distance to the surface) is calculated, and in a case where the difference is a preset threshold or below, it is determined that the voxel is visible from the camera. In a case where the voxel is determined as being visible, the pixel value of the coordinates (Xi, Yi) in the captured image from the image capture device **2** corresponding to the camera is used as the color of the voxel. In a case where the voxel is determined as being visible from a plurality of cameras, a pixel value is obtained from the texture data on the foreground image from each of the captured images from the image capture devices **2**, and for example, their average value is used as the color of the voxel. However, the color calculation method is not limited to this. For example, instead of using the average value, a pixel value in a captured image obtained from the image capture device **2** closest to the virtual viewpoint may be used. By repeating the same processing for all the voxels, colors can be assigned to all the voxels forming the three-dimensional shape data.

[0067] Although this example describes an example where the virtual viewpoint image generation unit **12** determines from which camera each three-dimensional point is visible, the present disclosure is not limited to this. In a case where the shape estimation device **3** has visibility information representing from which camera each three-dimensional point is visible, the virtual viewpoint image generation unit **12** may obtain the visibility information from the shape estimation device **3** and perform processing using the information thus obtained.

[0068] Next, a description is given of a method for generating a background viewpoint image of a virtual viewpoint image. To generate a background virtual viewpoint image, three-dimensional shape data on a background such as a stadium is obtained. As three-dimensional shape data on a background, a CG model of the stadium or the like is created in advance, and the CG model saved in the system is used. Vectors normal to the respective surfaces forming the CG model are compared to directional vectors of the cameras forming the image capture devices **2** to calculate the image capture device **2** having the surfaces within its angle of view and most directly facing them. Then, vertex coordinates of the surfaces are projected onto this image capture device **2**, and texture images to be attached to the surfaces are generated and rendered using an existing texture mapping method. A background virtual viewpoint image is thus generated. A virtual viewpoint image is generated by superimposing (combining) the foreground virtual viewpoint image on (with) the background virtual viewpoint image for the virtual viewpoint image thus generated.

[0069] The description of the flowchart in FIG. **5**A is continued. In S**504**, the teaching data generation unit **131** generates teaching data for learning of a neural network for repairing a virtual viewpoint image. Here, teaching data having a pair of input data and answer data is generated, the input data being the virtual viewpoint image corresponding to the viewpoint position of an actual camera, which has been generated in S**503**, the answer data being the captured image from the actual camera corresponding to the virtual viewpoint position, which has been obtained in S**501**. The color information in the virtual viewpoint image corresponding to the viewpoint position of the actual camera is equal to the image from the actual camera used for the shooting. Thus, this is because the virtual viewpoint image and the actual camera image are ideally equal to each other in a case where the position and attitude of a virtual viewpoint are the same as the position and attitude of the actual camera. Note that in S**504**, to have as many pieces of teaching data as needed for learning, data augmentation may be performed concomitantly. Examples of data augmentation methods to employ include methods employing the following processing. Specifically, as an example, there is a method employing, on a virtual viewpoint image which is input data and an actual camera image which is answer data corresponding thereto, processing of randomly cutting out the same corresponding image portion region (however, the cut image size is fixed) and processing of performing mirror inversion.

[0070] In S**505**, the repair learning unit **132** performs learning model (neural network) learning using the teaching data generated in S**504**. For example, the learning model is learned so that in response to an input of any given virtual viewpoint image, a virtual viewpoint image removed of or reduced in jelly noise can be generated as an output. A learned model obtained by the learning is outputted to the repair unit **133**. Note that as a loss function used in the neural network learning, mean square error is used to measure the fidelity of the input with respect to the answer. Also, Adam is used as a method for optimizing neural network parameters to minimize the loss function. Also, as the architecture of the neural network, an architecture equivalent to the architecture known as U-Net is used. U-Net is a network architecture for performing processing while performing multiresolution analysis on images, and is characteristically robust with respect to the scale of image features. For this reason, it is possible to handle jelly noise of various sizes, and it is expected to be effective for the virtual viewpoint image repair here. This is the processing performed in the learning phase.

[0071] Next, a description is given of inference processing for actually repairing a virtual viewpoint image using a learned neural network model.

[0072] FIG. **5**B is a diagram showing an example flowchart of inference processing for repairing a virtual viewpoint image using a learned neural network model. The camera information obtaining processing in S**501** and the shape estimation information obtaining processing in S**502** in FIG. **5**B are the same as those in FIG. **5**A and are therefore not described here.

[0073] After S**502**, in S**513**, the virtual viewpoint image generation unit **12** generates a virtual viewpoint image from any given viewpoint position. The method for generation the virtual viewpoint image is the same as the method described with S**504** in FIG. **5**A. In the inference phase, a virtual viewpoint image from any given viewpoint position is generated. The virtual viewpoint image generated is outputted to the repair unit **133** to be inputted to the learned model.

[0074] In S**514**, the repair unit **133** inputs the virtual viewpoint image generated in S**513** to the learned model learned in S**505** and thereby repairs the virtual viewpoint image. Note that any given virtual viewpoint image is inputted here regardless of whether the virtual viewpoint image has jelly noise or not. The learning carried out in S**505** is performed based on teaching data generated in S**503**, and the teaching data also includes virtual viewpoint images without jelly noise. Thus, it is expected not to perform unnecessary repair in a case where a virtual viewpoint image without jelly noise is inputted. The virtual viewpoint image repaired by the repair unit **133** is outputted to the display device **4**.

[0075] Also, for example, a configuration may be employed in which the repair unit **133** repairs a virtual viewpoint image only in a case where the virtual viewpoint image has jelly noise. In this case, the image generation apparatus **1** may have a determination unit that determines whether a virtual viewpoint image contains jelly noise. The determination unit may be included in the virtual viewpoint image repair unit **13**. Then, a virtual viewpoint image outputted from the virtual viewpoint image generation unit **12** is inputted to the determination unit, and the determination unit determines whether the inputted virtual viewpoint image contains jelly noise. Then, in a case where the virtual viewpoint image contains jelly noise, the virtual viewpoint image is outputted to the repair unit **133** and undergoes repair processing in the repair unit **133**. Meanwhile, in a case where the virtual viewpoint image does not contain jelly noise, the virtual viewpoint image bypasses the repair unit **133** and is outputted from the determination unit to the display device **4**.

[0076] Alternatively, a configuration may be employed in which a virtual viewpoint image generated by the virtual viewpoint image generation unit **12** is outputted to the virtual viewpoint image repair unit **13** for an event where jelly noise is likely to occur. For example, this configuration is employed for an event such as rugby where objects tend to get very close to each other, because a region uncapturable by any of the image capture devices tends to be generated, making it likely for jelly noise to occur. Meanwhile, for an event where subjects are unlikely to get very close to each other, a virtual viewpoint image generated by the virtual viewpoint image generation unit **12** may bypass the virtual viewpoint image repair unit **13** and be outputted directly to the display device **4**. To achieve this configuration, for example, the destination to which the virtual viewpoint image generation unit **12** outputs a virtual viewpoint image may be switched automatically between the virtual viewpoint image repair unit **13** and the display device **4** based on event information. Alternatively, besides the event information, the output destination may be switched based on information indicating a change in a possibility of jelly noise occurrence, such as the closeness of subjects. Also, the image processing apparatus **1** may be configured such that the output destination is switched according to a user operation or settings.

[0077] Also, although the above learning uses teaching data formed by a pair of input data and answer data on the same event held in the same venue, the present disclosure is not limited to this. Specifically, learning may be performed using teaching data including pairs of input data and answer data that are pairs of captured images captured in various events held in a plurality of different venues and virtual viewpoint images generated thereon. For example, teaching data A may be generated based on image capture of a rugby game held in a venue A, and teaching data B may be generated based on image capture of a rugby game held in a venue B. The learning of the repair learning unit **132** may be performed using teaching data including the teaching data A and the teaching data B. Further, in addition to the teaching data A and the teaching data B, the teaching data may include teaching data C generated based on image capture of a soccer game held in a venue C, and the learning by the repair learning unit **132** may be performed using such teaching data. Also, data suitable for learning may be selected from teaching data based on information on an event or the like or user settings, and learning may be performed based on the selected teaching data.

[0078] A configuration may be employed in which jelly noise and other noise are identified in a virtual viewpoint image outputted from the virtual viewpoint image generation unit **12**, e.g., automatically or according to user settings, and the virtual viewpoint image in which noise is identified is inputted to the teaching data generation unit.

[0079] As thus described, according to the present embodiment, jelly noise generated due to low-accuracy shape estimation results can be removed from a virtual viewpoint image by the after the fact processing. As a result, degradation of the image quality of a virtual viewpoint image can be reduced.

Second Embodiment

[0080] In the present embodiment, processing to detect a region with jelly noise in a virtual viewpoint image and to repair the detected region is learned, divided into two neutral networks: one for detection and one for repair. Specifically, a first model for detection and a second model for repair are learned. Then, in the example to be described, these learned models are combined to have the neural networks infer repair results.

<System Configuration>

[0081] FIG. **7** is a diagram showing the configuration of an image processing system of the present embodiment. The image processing system of the present embodiment includes an image generation apparatus **7** in place of the image generation apparatus **1** described in the first embodiment. As shown in FIG. **7**, the image generation apparatus **7** is connected to the image capture devices **2**, the shape estimation device **3**, and the display device **4** in a daisy chain or via a predetermined network. The configurations of the image capture devices **2**, the shape estimation device **3**, and the display device **4** are the same as those in the first embodiment. The following omits descriptions for configurations that are the same as those in the first embodiment.

[0082] The image generation apparatus **7** has the camera information obtainment unit **11**, the virtual viewpoint image generation unit **12**, and a virtual viewpoint image repair unit **73**. Compared to the first embodiment, the function and operation of the virtual viewpoint image repair unit **73** are different.

[0083] The virtual viewpoint image repair unit **73** detects which region has jelly noise in a virtual viewpoint image generated by the virtual viewpoint image generation unit **12**, and repairs the detected jelly noise region. This process is described using FIGS. **8**A and **8**B.

[0084] FIGS. **8**A and **8**B are diagrams illustrating a jelly noise map. FIG. **8**A is a diagram showing a jelly noise map which is an image representing jelly noise regions, which is obtained by inputting a virtual viewpoint image like the one represented by the image **221** in FIG. **2**C. FIG. **8**B is a diagram illustrating a virtual viewpoint image in which the jelly noise regions shown in FIG. **8**A have been repaired.

[0085] FIG. **8**A shows a jelly noise map **801** for the example of the image **221** in FIG. **2**C. Regions **805**, **806**, **807** in FIG. **8**A are pixel regions corresponding to the regions **225**, **226**, **227** in the image **221** in FIG. **2**C that are observed as jelly noise, respectively. An image **611** in FIG. **8**B is an example virtual viewpoint image in which the jelly noise regions have been repaired based on the jelly noise map **801** in FIG. **8**A. Regions **812**, **813**, **814** in FIG. **8**B are example image regions corresponding to the objects **202**, **203**, **204** in FIG. **2**A, respectively. In the present embodiment, jelly noise regions are detected, and the detected regions are targeted for repair, so that other image regions are not changed unnecessarily; thus, it is expected that the image quality of the virtual viewpoint image is improved stably.

[0086] The present embodiment assumes that the processing to detect a jelly noise region and to repair the jelly noise region is learned by two separated neural networks, and these two learned models are combined to repair a virtual viewpoint image. The virtual viewpoint image repair unit **73** of the present embodiment has a noise detection teaching data generation unit **731**, a noise detection learning unit **732**, a noise detection unit **733**, a repair teaching data generation unit **734**, a repair learning unit **735**, and a region repair unit **736**.

[0087] The noise detection teaching data generation unit **731** generates teaching data having, for example, the following pair. Specifically, the noise detection teaching data generation unit **731** generates teaching data formed by input data and answer data, the input data being a virtual viewpoint image generated by the virtual viewpoint image generation unit **12**, the answer data being a difference region between the virtual viewpoint image and a captured image from a camera having the corresponding viewpoint obtainable from the camera information obtainment unit **11**. Note that as the camera captured image used as the answer data, an image obtained by actually shooting a real space may be used, or an image generated by interpolation of captured images from two actual cameras may be used. Also, a camera simulation image obtained in a virtual three-dimensional space created by CG may be used. Constraints for these cases are the same as those in the example described in the first embodiment.

[0088] The noise detection learning unit **732** defines a loss function of the input with respect to the answer based on the teaching data generated by the noise detection teaching data generation unit **731**. Then, neural network parameters are repeatedly optimized so that the loss function can be minimized or maximized, and learning is thus conducted. Then, the model obtained by the learning is outputted to the noise detection unit **733**.

[0089] FIGS. **9**A and **9**B are diagrams illustrating an overview of a learning model for detecting jelly noise regions. FIG. **9**A shows an example of learning processing performed by the noise detection learning unit **732**. Learning is performed using teaching data formed by input data and answer data, the input data being a virtual viewpoint image P**1** corresponding to the viewpoint position of an actual camera C**1**, the answer data being a difference region between the virtual viewpoint image P**1** and a captured image captured by the actual camera C**1**. Then, learning is repeated to minimize or maximize an offset amount L between the input data and the answer data. Although an actual camera at one viewpoint is taken as an example here, learning is performed repeatedly using teaching data at the corresponding viewpoints of the image capture devices **2** forming the image processing system.

[0090] Note that the noise detection learning unit **732** may include an error detecting unit and an updating unit, and their functions are the same as those included in the repair learning unit **132** described in the first embodiment. Also, the present embodiment assumes that deep learning, which itself generates feature amounts and connection weighting coefficients for learning, is performed using a neural network. Note that as the network structure of a neural network used, any method may be employed as long as an input to and an output from the network are image data and the relation between the input and the output can be learned sufficiently.

[0091] The noise detection unit **733** inputs a virtual viewpoint image to a learned model obtained by the noise detection learning unit **732** and thereby detects which region in the virtual viewpoint image has jelly noise. The jelly noise region detected here may be outputted to the repair teaching data generation unit **734** and the region repair unit **736** after being converted to an image format which is called a jelly noise map and has the same size as the inputted virtual viewpoint image. Note that the learning may be performed so that the jelly noise map itself is outputted from the noise detection learning unit **732**. Also, the virtual viewpoint image given as an input may also be outputted to the repair teaching data generation unit **734** and the region repair unit **736**. In the repair learning phase, the virtual viewpoint image given as an input and the jelly noise map obtained from the neural network are outputted to the region repair unit **736**. In the repair inference phase, the virtual viewpoint image given as an input and the jelly noise map obtained from the neural network are outputted to the region repair unit **736**.

[0092] FIG. **9**B shows an example of jelly noise region detection processing (inference processing) performed by the noise detection unit **733**. As a result of inputting, as input data, a virtual viewpoint image P**2** at any given virtual viewpoint to the learned model obtained by the noise detection learning unit **732**, a jelly noise region R**2** in the virtual viewpoint image P**2** is detected. Then, the jelly noise region R**2** is converted to a jelly noise map M**2** having the same size as the virtual viewpoint image P**2**.

[0093] The repair teaching data generation unit **734** generates teaching data formed by a pair of input data and answer data, the input data being the virtual viewpoint image and the jelly noise map obtained from the noise detection unit **733**, the answer data being a captured image from a camera having the corresponding viewpoint obtainable from the camera information obtainment unit **11**. Note that like in the example described with the noise detection teaching data generation unit **731**, as the camera captured image used as answer data, an image obtained by actually shooting a real space may be used, or an image generated by interpolation of captured images from two actual cameras may be used. Also, a camera simulation image obtained in a virtual three-dimensional space created by CG (computer graphics) may be used. Constraints for these cases are the same as those in the example described in the first embodiment.

[0094] The repair learning unit **735** defines a loss function of the input with respect to the answer based on the teaching data generated by the repair teaching data generation unit **734**. Then, neural network parameters are repeatedly optimized so that the loss function can be minimized or maximized, and the learning is thus conducted. Then, the model obtained by the learning is outputted to the region repair unit **736**.

[0095] FIGS. **10**A and **10**B are diagrams illustrating an overview of a learning model for repairing a jelly noise region in a virtual viewpoint image. FIG. **10**A shows an example of learning processing performed by the repair learning unit **735**. Learning is performed using teaching data formed by input data and answer data, the input data being a virtual viewpoint image P**1** corresponding to the viewpoint position of an actual camera C**1** and a jelly noise map M**1** corresponding to the virtual viewpoint image P**1**, the answer data being a captured image captured by the actual camera C**1**. Then, learning is repeated to minimize or maximize an offset amount L between the input data and the answer data. Although an actual camera at one viewpoint is taken as an example here, learning is performed repeatedly using teaching data at the corresponding viewpoints of the image capture devices **2** forming the image processing system.

[0096] Note that the repair learning unit **735** may include an error detecting unit and an updating unit, and their functions are the same as those included in the repair learning unit **132** described in the first embodiment. Also, the present embodiment assumes that deep learning, which itself generates feature amounts and connection weighting coefficients for learning, is performed using a neural network. Note that as the network structure of a neural network used, any method may be employed as long as an input to and an output from the network are image data and the relation between the input and the output can be learned sufficiently.

[0097] The region repair unit **736** inputs the jelly noise map and the virtual viewpoint image given from the noise detection unit **733** to the learned model obtained by the repair learning unit **735** and thereby repairs the virtual viewpoint image. The repaired virtual viewpoint image is outputted to the display device **4**.

[0098] FIG. **10**B shows an example of jelly noise region repair processing (inference processing) performed by the region repair unit **736**. A virtual viewpoint image P**2** from any given virtual viewpoint and a jelly noise map M**2** corresponding to the virtual viewpoint image P**2** are inputted as input data to the learned model obtained by the repair learning unit **735**. Then, a repaired virtual viewpoint image in which the jelly noise region R**2** in the virtual viewpoint image P**2** has been repaired is outputted from the learned model.

<Processing Flow>

[0099] FIGS. **11** and **12** are flowcharts showing an example of processing performed by the image generation apparatus **7** of the present embodiment. Using the flowcharts shown in FIGS. **11** and **12**, a description is given of processing performed by the image generation apparatus **7** of the present embodiment. Note that steps denoted by the same numbers as those in the flowchart in FIGS. **5**A and **5**B are the same as the steps described in the first embodiment and are therefore not described here.

[0100] First, a flowchart for processing for learning of a neural network for detecting a jelly noise region in a virtual viewpoint image is described using FIG. **11**A. After the processing in S**501** and S**502**, processing in S**1103** is performed.

[0101] In S**1103**, the virtual viewpoint image generation unit **12** generates a group of virtual viewpoint images corresponding to the positions of the actual cameras. The group of virtual viewpoint images thus generated are outputted to the noise detection teaching data generation unit **731**. The virtual viewpoint images generated in S**1103** are used as input data for neural network learning. Note that unlike the first embodiment, in S**1103**, the group of virtual viewpoint images outputted to the noise detection teaching data generation unit **731** may be only virtual viewpoint images containing jelly noise or may include virtual viewpoint images containing no jelly noise at a rate of approximately 1%. By predominantly using scenes in which jelly noise occurs as input data for learning, the characteristics of a jelly noise region can be learned predominantly. Also, by also adding a small number of virtual viewpoint images without jelly noise instead of using virtual viewpoint images all containing jelly noise, it is expected to improve the robustness of the learned model.

[0102] In S**1104**, the noise detection teaching data generation unit **731** calculates a difference image between a captured image from an actual camera obtained in S**501** and the virtual viewpoint image generated in S**1103** corresponding to the viewpoint position of this actual camera. Note that this difference image is a binary image such that each pixel of a foreground region has 1 as its pixel value in a case where the absolute value of the difference between the two images is a threshold or greater and has 0 otherwise. Note that all the pixels of a background image have 0 as their pixel values. In other words, pixels whose difference between two images is the threshold or greater are detected as jelly noise. Here, the threshold is an allowable value of whether to detect the pixel as jelly noise, and any value can be set depending on how much difference to allow. In the present embodiment, as an example, the threshold is set to 5. Note that in the example described in the present embodiment, a difference image between a virtual viewpoint image and a captured image from the corresponding viewpoint is used as answer data on a jelly noise region, but in S**1104**, it is only necessary to be able to obtain data (image data) to be used as answer data. In a different example, a weighted image based on the visibility of a group of three-dimensional points forming a subject from each camera may be obtained, or a mask image having a jelly noise region manually specified by a user may be obtained. Note that a weighted image based on the visibility from each camera is a weighted image generated such that a pixel which is projection of a three-dimensional point of note onto the camera has a weight of 1 in a case where the three-dimensional point is visible from the camera and has a weight of 0 in a case where the three-dimensional point is invisible from the camera. This is because jelly noise often occurs at a region invisible from the group of cameras used for shooting, and it is therefore expected that a jelly noise region is detected inside the weighted image. Note that in a case where a user manually specifies a jelly noise region, a jelly noise map may be created from the start based only on virtual viewpoint images. Also, a corrected image may be used such that an image representing a jelly noise map created by the above method is corrected only in a region with excess or deficiency of jelly noise. In a case where a user manually specifies a jelly noise region, a step for specifying a jelly noise region is additionally provided.

[0103] In S**1105**, the noise detection teaching data generation unit **731** generates teaching data for learning of a neural network for detecting jelly noise in a virtual viewpoint image. Note that here, teaching data formed by a pair of input data and answer data is generated, the input data being the virtual viewpoint image generated in S**1103**, the answer data being the difference image calculated in S**1104**. Since the color information for a virtual viewpoint image is equal to that for an image from an actual camera used for the shooting, the virtual viewpoint image and the actual camera image are ideally equal to each other in a case where the position and attitude of the virtual viewpoint and the position and attitude of the actual camera are the same. Thus, this is because the difference image is expected to have a jelly noise region emerging therefrom. Note that in S**1105**, to have as many pieces of teaching data as needed for learning, data augmentation may be performed concomitantly. Examples of data augmentation methods to employ include methods employing the following processing. Specifically, there is a method employing, on a virtual viewpoint image which is input data and a difference image which is answer data corresponding thereto, processing of randomly cutting the same corresponding image portion region (however, the cut image size is fixed) and processing performing mirror inversion.

[0104] In S**1106**, the noise detection learning unit **732** performs neural network learning using the teaching data generated in S**1105**. More specifically, the noise detection learning unit **732** performs neural network learning so that a jelly noise map which is an image representing a jelly noise region can be generated as an output in response to input of any given virtual viewpoint image. The learned model obtained by the learning is outputted to the noise detection unit **733**.

[0105] In the present embodiment, a jelly noise map which is an image representing a jelly noise region and obtained as an output of the learned model is assumed to be such that each pixel has a pixel value indicating whether it is jelly noise, i.e., **0** or **1** (binary). Thus, the jelly noise detection can be interpreted as the labeling problem, and thus, cross-entropy loss used for evaluation of whether the label is correct is used as a loss function for use in neural network learning. Also, the stochastic gradient method is used as a method for optimizing neural network parameters to minimize the loss function. Also, as the architecture of the neural network, an architecture equivalent to the architecture used in SegNet is used, SegNet being known as being capable of highly-accurate segmentation. Although a jelly noise map is binary in the processing performed in the present embodiment, the processing may be performed handling a jelly noise map as multilevel. In that case, the labels may be divided into multilevel labels, or a pixel value may be regarded not as a label but as the likelihood of jelly noise so that a probability, not a label value, may be outputted as an output from the neural network for each pixel.

[0106] Note that a user may add processing to the jelly noise map. For example, in a later jelly noise repair NN, a user may identify noise that the user wants repaired at the same time, and annotating processing may be performed on an image region of the noise thus identified. All that is needed is to make the pixel values of the region thus annotated be the same as the pixel values of the jelly noise region. Also, a user may identify noise that the user wants repaired at the same time, and a combined map generated from a jelly noise map and a map including the region of the identified noise may be used as a jelly noise map, the combined map having, as noise, a region included as noise in either of the maps.

[0107] Next, a description is given of a flowchart shown in FIG. **11**B for processing of learning of a neural network for repairing a jelly noise region in a virtual viewpoint image. The processing steps S**501**, S**502**, S**1103** are the same as those shown in FIG. **11**A. After these processing steps, processing in S**1114** is performed.

[0108] In S**1114**, the noise detection unit **733** generates a jelly noise map by inputting the virtual viewpoint image corresponding to the actual camera position generated in S**1103** to the learned model obtained by the noise detection learning unit **732**. The generated jelly noise map is outputted to the repair teaching data generation unit **734**.

[0109] In S**1115**, the repair teaching data generation unit **734** generates teaching data for neural network learning for performing repair processing on the jelly noise region in the virtual viewpoint image. The teaching data generated here is formed by input data and answer data, the input data being the virtual viewpoint image generated in S**1103** and the jelly noise map generated in S**1114**, the answer data being the captured image from the actual camera corresponding to the position of the virtual viewpoint, obtained in S**501**. This is because, since the color information for a virtual viewpoint image is equal to that for an image from the actual camera used for the shooting, the virtual viewpoint image and the actual camera image are ideally equal to each other in a case where the position and attitude of the virtual viewpoint and the position and attitude of the actual camera are the same.

[0110] In S**1116**, the repair learning unit **735** performs neural network learning using the teaching data generated in S**1115**. More specifically, the repair learning unit **735** performs neural network learning so that a virtual viewpoint image in which the jelly noise region has been repaired can be generated as an output in response to input of any given virtual viewpoint image and a jelly noise map corresponding thereto. Note that the virtual viewpoint image and the jelly noise map that are given as an input are inputted to a single layer in the neural network, i.e., as a single multi-channel image integrating the virtual viewpoint image and the jelly noise map. The learned model obtained by the learning is outputted to the region repair unit **736**. Note that as a loss function used in the neural network learning, mean square error is used to measure the fidelity of the input with respect to the answer. Note, however, that error is calculated only for pixels forming a region determined as jelly noise in the jelly noise map. By calculating error only for pixels forming a jelly noise region, the image quality of a non-jelly-noise region can be left unaffected. Also, Adam is used as a method for optimizing neural network parameters to minimize the loss function. Also, as the architecture of the neural network, an architecture having a partial convolution layer in place of a convolution layer in U-Net employed in the first embodiment is used. The partial convolution layer gives the positions of pixels used for computation as a mask image, and thereby performs processing using only the values in the masked region. Thus, a partial convolution layer is suitable for image inpainting processing. A partial convolution layer is effective because the virtual viewpoint image repair in the present embodiment can be interpreted as inpainting processing of a jelly noise region.

[0111] Next, using FIG. **12**, a description is given of a flowchart of inference processing for actually repairing a virtual viewpoint image using the learned neural network models for jelly noise region detection and repair. In FIG. **12**, processing in S**1204** is performed after the processing in S**501**, S**502**, and S**513** described in the first embodiment.

[0112] In S**1204**, the noise detection unit **733** inputs a virtual viewpoint image generated in S**513** to the learned model obtainable from the noise detection learning unit **732** and generates a jelly noise map. Note that any given virtual viewpoint image is inputted here in order to detect whether jelly noise is contained. The jelly noise map generated is outputted to the region repair unit **736** to be inputted to the learned model.

[0113] In **51205**, the region repair unit **736** inputs the corresponding virtual viewpoint image given and the jelly noise map generated in S**1204** to the learned model learned in S**1116** and thereby repairs the virtual viewpoint image. Note that any given virtual viewpoint image is inputted regardless of the presence of jelly noise. This is because the learned model learned in S**1116** has been learned to repair only jelly noise regions, and unless a jelly noise region is detected in S**1205**, other regions are unaffected. As a result, jelly noise regions can be improved with side effects mitigated. The repaired virtual viewpoint image is outputted to the display device **4**.

[0114] Although any given virtual viewpoint image is inputted to the region repair unit **736** regardless of the presence of jelly noise in the processing in FIG. **12** in the example described, the present disclosure is not limited to this. In a case where there is no jelly noise region in the jelly noise map generated in S**1204**, the corresponding virtual viewpoint image may be not inputted to the region repair unit **736** to omit repair processing.

[0115] As thus described, the present embodiment can detect which region in a virtual viewpoint image has jelly noise which occurs due to low-accuracy shape estimation results and repair the jelly noise region based on the detection result. Thus, a virtual viewpoint image can be repaired with non-jelly-noise regions unaffected. As a result, it is possible to reduce degradation of the image quality of the virtual viewpoint image.

OTHER EMBODIMENTS

[0116] Although the above embodiments have described examples where the image generation apparatus includes both the learning unit and the repair unit or the detection unit (inference unit), the learning unit and the inference unit may be included in separate image generation apparatuses. For example, learning may be performed in a first image generation apparatus including the learning unit. Then, the learned model learned may be sent to a second image generation apparatus including the inference unit, and inference processing may be performed in the second image generation apparatus.

[0117] Also, in the above embodiments, the learned model may be created in a different environment (outside the image processing system in FIG. **1**), and noise repair may be performed by applying the learning results.

[0118] Also, although noise regions are corrected using machine learning in the above embodiments, the present disclosure is not limited to this. It is also possible to obtain a virtual viewpoint image removed of or reduced in noise by extracting a difference by comparison between a virtual viewpoint image from a predetermined virtual viewpoint and an image from an actual camera which has a viewpoint which is the same as or closest to the virtual viewpoint and correcting the virtual viewpoint image using the difference. In this processing, the comparison may be performed after performing projective transformations or the like to bring the actual camera image to or closer to the virtual viewpoint of the virtual viewpoint image to be compared with. Also, a virtual viewpoint image may be compared with an image obtained by appropriately blending a plurality of actual camera images (combining processing).

[0119] Embodiment(s) of the present disclosure can also be realized by a computer of a system or apparatus that reads out and executes computer executable instructions (e.g., one or more programs) recorded on a storage medium (which may also be referred to more fully as a 'non-transitory computer-readable storage medium') to perform the functions of one or more of the above-described embodiment(s) and/or that includes one or more circuits (e.g., application specific integrated circuit (ASIC)) for performing the functions of one or more of the above-described embodiment(s), and by a method performed by the computer of the system or apparatus by, for example, reading out and executing the computer executable instructions from the storage medium to perform the functions of one or more of the above-described embodiment(s) and/or controlling the one or more circuits to perform the functions of one or more of the above-described embodiment(s). The computer may comprise one or more processors (e.g., central processing unit (CPU), micro processing unit (MPU)) and may include a network of separate computers or separate processors to read out and execute the computer executable instructions. The computer executable instructions may be provided to the computer, for example, from a network or the storage medium. The storage medium may include, for example, one or more of a hard disk, a random-access memory (RAM), a read only memory (ROM), a storage of distributed computing systems, an optical disk (such as a compact disc (CD), digital versatile disc (DVD), or Blu-ray Disc (BD)™), a flash memory device, a memory card, and the like.

[0120] While the present disclosure has been described with reference to exemplary embodiments, it is to be understood that the disclosure is not limited to the disclosed exemplary embodiments. The scope of the following claims is to be accorded the broadest interpretation so as to encompass all such modifications and equivalent structures and functions.

1. An image processing apparatus comprising:
one or more memories storing instructions; and
one or more processors executing the instructions to:
obtain a virtual viewpoint image generated based on a plurality of captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints and three-dimensional shape data on the object; and
remove noise in the obtained virtual viewpoint image, the noise being generated due to accuracy of the three-dimensional shape data.

2. The image processing apparatus according to claim **1**, wherein
the noise is removed using a learned model learned based on teaching data having input data and answer data, the input data being a virtual viewpoint image from a viewpoint corresponding to a predetermined position and attitude, the answer data being an image corresponding to the predetermined position and attitude.

3. The image processing apparatus according to claim **2**, wherein
the predetermined position and attitude are a position and attitude of a predetermined image capture device, and
the image corresponding to the predetermined position and attitude is a captured image obtained by image capture by the predetermined image capture device.

4. The image processing apparatus according to claim **1**, wherein
the one or more processors further execute the instructions to
detect a region with the noise in the virtual viewpoint image and
repair the region with the noise in the detected virtual viewpoint image.

5. The image processing apparatus according to claim **4**, wherein
the detection is performed using a first model which is a learned model learned based on teaching data having input data and answer data, the input data being a virtual viewpoint image from a viewpoint correspond-

ing to a predetermined position and attitude, the answer data being a difference image indicating a difference between the virtual viewpoint image and an image corresponding to the predetermined position and attitude.

6. The image processing apparatus according to claim **5**, wherein

a corrected image obtained by a user correcting the difference image is used as the answer data instead of the difference image.

7. The image processing apparatus according to claim **4**, wherein

the detection is performed using a first model which is a learned model learned based on teaching data having input data and answer data, the input data being a virtual viewpoint image from a viewpoint corresponding to a predetermined position and attitude, the answer data being an image based on visibility of three-dimensional shape data forming an object in an image corresponding to the predetermined position and attitude, the object shown in the image.

8. The image processing apparatus according to claim **7**, wherein

a corrected image obtained by a user correcting the image based on visibility is used as the answer data instead of the image.

9. The image processing apparatus according to claim **4**, wherein

the detection is performed using a first model which is a learned model learned based on teaching data having input data and answer data, the input data being a virtual viewpoint image from a viewpoint corresponding to a predetermined position and attitude, the answer data being a mask image such that a user sets a noise region in an image corresponding to the predetermined position and attitude.

10. The image processing apparatus according to claim **9**, wherein

the predetermined position and attitude are a position and attitude of a predetermined image capture device, and

the image corresponding to the predetermined position and attitude is a captured image obtained by image capture by the predetermined image capture device.

11. The image processing apparatus according to claim **4**, wherein

the repair is performed using a second model which is a learned model learned based on teaching data having input data and answer data, the input data being a virtual viewpoint image from a viewpoint corresponding to a predetermined second position and attitude and the region in the detected virtual viewpoint image detected as having the noise occurring therein, the answer data being an image corresponding to the second position and attitude.

12. The image processing apparatus according to claim **11**, wherein

a region having noise occurring in the virtual viewpoint image is detected and the virtual viewpoint image and the detected region is outputted, and

the outputted virtual viewpoint image and the outputted region is inputted to the second model and thereby the noise in the region is repaired.

13. The image processing apparatus according to claim **11**, wherein

the second position and attitude are a position and attitude of a predetermined image capture device, and

the image corresponding to the second position and attitude is a captured image obtained by image capture by the predetermined image capture device.

14. An image processing apparatus comprising:

one or more memories storing instructions; and

one or more processors executing the instructions to:

obtain a virtual viewpoint image corresponding to a virtual viewpoint, the virtual viewpoint image being generated based on a plurality of captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints; and

correct the obtained virtual viewpoint image based on, among the plurality of captured images, at least a captured image captured by an image capture device that captures the object from a viewpoint corresponding to the virtual viewpoint.

15. An image processing apparatus comprising:

one or more memories storing instructions; and

one or more processors executing the instructions to:

obtain a virtual viewpoint image corresponding to a virtual viewpoint, the virtual viewpoint image being generated based on a plurality of captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints; and

correct the obtained virtual viewpoint image based on learning results based on teaching data having input data and answer data, the input data being a plurality of virtual viewpoint images corresponding to the plurality of viewpoints, the answer data being the plurality of captured images obtained by image capture by the plurality of image capture devices corresponding to the plurality of viewpoints.

16. An image processing method comprising:

obtaining a virtual viewpoint image generated based on a plurality of captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints and three-dimensional shape data on the object; and

removing noise in the virtual viewpoint image obtained by the obtaining, the noise being generated due to accuracy of the three-dimensional shape data.

17. An image processing method comprising:

obtaining a virtual viewpoint image corresponding to a virtual viewpoint, the virtual viewpoint image being generated based on a plurality of captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints; and

correcting the virtual viewpoint image obtained by the obtaining based on, among the plurality of captured images, at least a captured image captured by an image capture device that captures the object from a viewpoint corresponding to the virtual viewpoint.

18. An image processing method comprising:

obtaining a virtual viewpoint image corresponding to a virtual viewpoint, the virtual viewpoint image being generated based on a plurality of captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints; and correcting the virtual viewpoint image obtained by the obtaining based on learning results based on teaching data having input data and answer data, the input data being a plurality of virtual viewpoint images corresponding to the plurality of viewpoints, the answer data being the plurality of captured images obtained by image capture by the plurality of image capture devices corresponding to the plurality of viewpoints.

19. A method for generating a learned model, the method comprising:

obtaining a plurality of captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints and a plurality of virtual viewpoint images corresponding to the plurality of viewpoints, the plurality of virtual viewpoint images being generated based on the plurality of captured images; and

generating a learned model based on teaching data having input data and answer data, the input data being the virtual viewpoint images obtained by the obtaining, the answer data being the captured images obtained by the obtaining.

20. A non-transitory computer readable storage medium storing a program which causes a computer to execute

obtaining a virtual viewpoint image generated based on a plurality of captured images obtained by image capture of an object by a plurality of image capture devices from a plurality of viewpoints and three-dimensional shape data on the object; and

removing noise in the virtual viewpoint image obtained by the obtaining, the noise being generated due to accuracy of the three-dimensional shape data.

* * * * *